INFLUENCE OF ROADWAY CHARACTERISTICS IN THE MODELING OF THE FREQUENCY OF ROADWAY DEPARTURE CRASHES ON TWO-LANE TWO-WAY STATE ROADS

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERSITY OF HAWAIʻI AT MĀNOA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

CIVIL AND ENVIRONMENTAL ENGINEERING

APRIL 2019

By

Mohammadreza Hashemi

Dissertation Committee:

Adrián R. Archilla, Chairperson

Panos D. Prevedouros

Guohui Zhang

Lin Shen

Reza Ghorbani

Dedicated to my parents

for their love and support throughout my life. Thank you both for giving me the strength to chase my dreams.

Acknowledgments

Foremost, I would like to express my sincere appreciation to my advisor, Dr. Adrián R. Archilla, professor of the department of civil and environmental engineering, for his continuous support during my Ph.D. research, his patience, guidance, and immense knowledge.

Besides my advisor, I would like to thank the rest of my dissertation committee members: Dr. Panos D. Prevedouros, Dr. Guohui Zhang, Dr. Lin Shen, and Dr. Reza Ghorbani.

I sincerely acknowledge all the support offered by the Traffic, Materials Testing and Research, and Planning Branches of the Hawaii Department of Transportation (HDOT). Particularly I am very much thankful to Jan Higaki, Sean Hiraoka, and Wayne Kahawara *for their amazing support, patience, and their trust.*

Last but not least, I would like to express my deepest appreciation to my family and friends for their assistance and encouragement in every step of my life. The completion of my dissertation would not have been possible without their support.

ABSTRACT

According to the Federal Highway Administration (FHWA), Roadway Departure (RwD) crashes account for approximately 56 percent of traffic fatalities in the United States. Likewise, FHWA's statistics indicate that RwD crashes represent a high proportion (approximately 54 percent) of traffic fatalities in the State of Hawaii. Therefore, there is a need to study their contributing factors and to quantify their effects by developing statistical models that may provide better inferences to alleviate them.

Using ten years of crash data, this research explores the effect of roadway characteristics (e.g., traffic, geometry, etc.) in the modeling of the frequency of RwD crashes on Two-Lane Two-Way (TLTW) state roads in the State of Hawaii. Specifically, the study concentrates on the effects of segment length, roadway directional attributes, and the general geometric environment of the analysis segment. These factors are evaluated with various Generalized Linear Models (GLM) such as the negative binomial regression, zero-inflated negative binomial, and mixed-effects negative binomial regression model.

The results show that segment length affects the model's estimations (total number of statistically significant parameters and their estimated values) and present the development of recommendations about an appropriate segment length for modeling the frequency of RwD crashes. Also, it confirms that the consideration of directional analysis improves the quality of the models in two ways: firstly, by assigning the head-on crashes based to the direction of the vehicles causing the crashes, and secondly by identifying the contributing factors based on the direction of vehicles causing the crashes. Moreover, the results indicate that the general geometric environment of the roadway portion where the

segment was located affects the frequency of RwD crashes, which means that, for example, for two similar segments, the frequency of RwD crashes are not equal if one is located on a winding road and the other segment is located right after a tangent road. This finding is in accordance with design consistency practices.

Another benefit of this study is the development of robust and realistic crash frequency models for the state of Hawaii as well as the improvement of the identification of the RwD crashes' contributing factors. In practice, decision-makers may consider the results to prioritize the location and type of countermeasures to mitigate RwD on TLTW state roads in Hawaii effectively.

Other unique features of the estimated models include: 1) using an estimate of mean friction demand as an independent variable, 2) capturing the different effects of upgrades and downgrades, and 3) using Annual Average Daily Traffic (AADT) both as a measure of exposure and separately as an independent variable affecting the rate of RwD crashes.

Finally, a new approach consistent with the probabilistic nature of the estimated generalized regression models is introduced for evaluation of their goodness of fit. Its use is suggested as a complementary tool in the typical evaluation of generalized regression models.

Keywords: Roadway Departure Crashes, Roadway Characteristics, Geometry Design, Generalized Linear Modeling

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Background

Highway fatalities and severe injuries in the United States continue to be a major national concern. According to the Federal Highway Administration (FHWA), although there has been a downward trend in both the fatal crash rates and the total number of traffic fatalities, the fact that there are still 34,674 average annual fatalities (2007-2014) indicates that much work is still required *(1)*. A significant contributor to this number is Roadway Departure (RwD) crashes, which account for approximately 56 percent of traffic fatalities *(1)*.

FHWA published a strategic plan for Roadway Departure (RwD) crashes and defined a RwD crash as a non-intersection crash in which a vehicle crosses an edge line, a centerline, or otherwise leaves the traveled way *(1)*. Therefore, RwD crashes include both run-off the road (ROR) and head-on crashes. According to the American Association of State Highway and Transportation Officials (AASHTO), ROR crashes involve vehicles that leave the travel lane and encroach onto the shoulder and beyond and hit one or more of any number of natural or artificial objects, such as bridge walls, poles, embankments, guardrails, parked vehicles, and trees *(2)*. Also, a head-on crash typically occurs when a vehicle crosses a centerline or a median and crashes into an approaching vehicle. Head-on crashes occur as a result of a driver's negligent actions as with ROR encroachments or deliberate actions (e.g., executing a passing maneuver on a two-lane road) *(3)*.

Traffic fatalities have received significant attention on different national acts; whose goals include, among others, improving the safety of roads. The Moving Ahead for Progress in the 21st Century Act (MAP-21) went into effect on October 1, 2012. It continued the Highway Safety Improvement Program (HSIP) as a core Federal-aid program. The goal of the HSIP is to achieve a significant reduction in traffic fatalities and serious injuries on all public roads. The HSIP requires a data-driven, strategic approach to improve highway safety on all public roads. A major component of the HSIP is a Strategic Highway Safety Plan (SHSP) as developed by the American Association of State Highway and Transportation Officials (AASHTO) with the assistance of the FHWA, the National Highway Traffic Safety Administration (NHTSA), and the Transportation Research Board Committee on Transportation Safety Management. The SHSP is a statewide-coordinated safety plan. It identifies strategies in 22 State's key safety needs and guides investment decisions towards strategies and countermeasures with the most potential to save lives and prevent injuries *(4)*. Some of the SHSP strategies explicitly address RwD crashes. Therefore, focusing on reducing the total number of RwD crashes will contribute towards achieving the national goals in accordance with the SHSP and the HSIP.

Moreover, AASHTO has embraced a "Towards Zero Deaths" vision and a goal of cutting fatalities in half by 2030. To accomplish this goal, the total number of fatalities would have to be reduced by approximately 1,000 per year nationally. A reduction in RwD crashes has great potential to help achieve this goal. Hence, the FHWA established a strategic plan to provide a common vision to address RwD crashes *(1)*.

- *Vision* – Pursue a proactive approach that will lead "Toward Zero Deaths and Serious Injuries" involving RwD crashes.
- *Mission* – Exercise leadership in the highway community to reduce the risk of RwD fatal and severe injury crashes from occurring.
- *Goal* - Reduce national RwD fatalities by a minimum of 500 per year from the existing 17,000 per year to 8,500 per year by 2030.

Therefore, studying RwD crashes is an essential step towards fulfilling the vision of zero deaths and serious injuries as well.

The statistics in the State of Hawaii indicate that the majority of traffic fatalities (approximately 54 percent) are related to RwD crashes *(5)*. Meanwhile, studying crashes by considering roadway characteristics (e.g., traffic, geometry, etc.) is in its early stages in the State of Hawaii. Using analytical methods to study the contributing factors of RwD crashes should result in more informed decisions for selecting the locations and types of safety projects.

To the best of the writer's knowledge, no study resulting in the development of a crash frequency[1] model has been conducted in the State of Hawaii. The Highway Safety Manual (HSM) includes some crash frequency models *(6)*; however, its use presents several challenges for Hawaii. First, the models in the HSM have been calibrated with other

[1] Also known as Safety Performance Function (SPF) in Highway Safety Manual (HSM). A crash frequency model identifies risk factors and quantifies their effects on the frequency of crashes.

states' data and may not reflect the reality in the State of Hawaii. Most of those models are very coarse, with only a few explanatory variables such as Average Annual Daily Traffic (AADT) and length to predict the crashes; however, it is intuitively clear that many other factors may affect the frequency of crashes. Also, those models do not include other roadways characteristics, which makes it difficult to decide on any further safety improvements. Although the national Crash Modification Factor (CMF) clearinghouse repository can be used for this purpose, its applicability to Hawaii is unclear.

Although the number of Hawaii's RwD crashes that occur on Two-Lane Two-Way (TLTW) state roads represent only about 28 percent of all RwD crashes, approximately 40 percent of all RwD crashes' fatalities occurred on these roads despite their substantially lower exposure because of the lower traffic volumes they carry. Thus, RwD crashes in TLTW state roads represent an important subset of all RwD crashes worth studying to reduce the number of fatalities, injuries, and property damage in the state. There is currently a need to understand the factors that contribute to RwD crashes in Hawaii, to model their frequency and severity, and to identify the roadway characteristics that contribute to a higher number of RwD crashes. This research aims to provide additional insights toward understanding the contributing factors of RwD crashes on TLTW state roads by developing robust crash frequency models. These models are helpful to the identification of segments with a higher risk of RwD crashes. They may also be useful for decision-makers to select safety countermeasures with more insight to alleviate the RwD crashes.

It worth mentioning that the scope of this dissertation is to model the frequency of RwD crashes in TLTW state roads based on roadway characteristics, and it does not model

the severity of crashes that entails other information such as socioeconomic characteristics, vehicle characteristics, and weather data. Also, the analyses are based on all the RwD crashes without regard for the severity. Generally, there are very few fatalities per year in Hawaii (see section 3.3 for more details), so the available data is not adequate for modeling the frequencies by severity type. The models are limited to TLTW state roads. Multilane highways and freeways are outside the scope of this study.

1.2 Problem statement

The extensive literature on crash frequency modeling is mostly focused on the statistical approaches used for estimation of model parameters, but very little is found in terms of the methods used to synthesize the data before developing a statistical model. However, regardless of the statistical estimation approach used, the estimated model parameters may be affected substantially by data generation processes such as selected road segment lengths or the type of attributes selected for modeling each segment that will be fully discussed in next sections.

1.2.1 Roadway segmentation

Typically, researchers employ two approaches to select the roadway segments used for developing statistical models. One approach involves splitting the roadways into fixed-length roadway segments. Another approach consists of splitting roadways into homogeneous segments based on the values of a select group of variables such as lane width, shoulder width, and median type. This approach defines a new segment if any of those variables changes along the roads.

While the use of homogeneous segments is theoretically appealing, the attributes defining homogeneity are usually selected for convenience rather than how they affect crash frequencies. For example, careful reviews of the Hawaii Department of Transportation (HDOT) video log (with a frame approximately every 0.004 mile) as well as of the data for roadway geometry and other assets collected by Light Detection and Ranging (LiDAR) equipment on TLTW state roads make it clear that even in supposedly "homogeneous segments" there are considerable variations of several of the roadway attributes not used for the segmentation. This is to be expected since the selection of a few variables used to split roadways into homogenous segments is somewhat arbitrary. Furthermore, because of the typically higher precision and higher data density produced with new technologies, the data are not entirely consistent with traditional roadway inventory databases or straight-line diagrams of the road. The changes may be related to roadway safety improvements (e.g., installation of a very short median in part of the roads) or roadway geometry improvements (e.g., widening of the shoulder).

On the other hand, using all the possible roadway characteristics to develop nearly homogenous segments would result in too many tiny segments, particularly if the databases were collected with high precision equipment like LiDAR. This is problematic for modeling the frequency of crashes since the number of segments with no crashes is increased and the maximum number of observed crashes on each segment may be considerably limited, which defeats the goal of modeling counts. The third potential problem is that segments of differing length may be problematic if the hazard rate is not constant. The use of fixed length segments tends to avoid this issue.

Moreover, some variables commonly used to define homogeneous segments such as lane width, the total number of lanes, and median type, are either not relevant for TLTW roadways or have little variation, and thus add little to a meaningful segmentation on this type of road. Of course, roads classified as TLTW roads may have short segments with two-way left turning lanes and with short medians, but for most of the network, these variables are not very useful for defining homogeneous segments. In addition, since the location of crashes as reported by the police are not usually exact, a homogenous segmentation approach that includes very tiny segments may be more prone to the assignment of crashes to the wrong segment. Likewise, if a variable such as a curvature were used to define homogeneous segments, the geolocation of a curve related crash may be identified right before or after the curve (although the crash may have indeed occurred on a straight segment, the adjacent curve may have been a major contributor). Thus, even if the frequencies for the small segments were generally acceptable for modeling purposes, these types of issues may obscure the actual causes of the crashes.

On the other hand, little to no justification is usually given for the selection of a fixed-length for segments other than that the data is only available for a given fixed length. In fact, it is not easy to process the data pertaining to the roadway characteristics (e.g., traffic, geometry, and inventory databases) for different segment lengths especially when the database has many features with different types of variables (i.e., continuous variables and categorical variables), or the network size (the total length of roads) is considerable. Meanwhile, no study was found in the literature that investigates the sensitivity of estimated crash frequency models by the segment length. Therefore, this dissertation

explores this issue by generating scripts that automate[1] the computation of the roadway characteristics for every given length.

In general, selecting an appropriate fixed-length represents a trade-off between the extreme of making the sections too long to capture the effects of relevant roadway features without them being washed out by substantial averaging and the other extreme where the segments are so short that relevant characteristics of the roadway environment are not captured adequately.

A possible advantage of studying the crashes using the fixed-length segmentation is its compatibility with the systemic analysis recommended by FHWA. In general, crashes are random events, and they fluctuate over time. FHWA published a report to promote the systemic analysis of crashes rather than the traditional black spots approach *(7)*. In the systemic approach, the objective is to reduce the total number of crashes preferably using low-cost countermeasures. In this regard, using homogenous segments may produce results similar to the traditional black spot approach if the high crash locations in the datasets represent unusual conditions unlikely to be repeated. Therefore, using an appropriate segment length (with more repeatable frequencies for a given set of characteristics) may be more consistent with a systemic approach and low-cost safety improvements. Also, presenting the results in terms of segments with a fixed length may facilitate the economic evaluation of safety projects, especially for low-cost safety improvement countermeasures

[1] For this reason, several scripts were developed using R programming.

such as edge line rumble strips, centerline rumble strips, pavement raised markers and chevron signs.

1.2.2 Roadway direction

The attributes selected for modeling crash frequencies on each segment are usually defined without regard of the direction of travel (i.e., it is assumed that the roadway characteristics are equal in both directions), and they are based on measurements strictly within the segment. However, it is possible that in some situations directional attributes may be important when the roadway characteristics (e.g., traffic, geometry, and inventory databases) have different values in each direction. Notably, it may play a crucial role in modeling the frequency of crashes on TLTW roads. For example, the shoulder width may differ widely from one direction to another. Likewise, for segments with steep grades, the effect of the grade could differ substantially for each direction of travel. Figure 1 shows two vehicles traveling in opposite directions on a grade. It is well known that because the component of the acceleration of gravity acts to accelerate vehicles traveling downgrade and decelerate vehicles traveling upgrade, maintaining the vehicle's speed below a safe value on steep downgrades, particularly for heavy vehicles, is more challenging than on steep upgrades. Other examples where conditions may differ by direction are the total length of guardrails in each direction, shoulder width, shoulder type, rutting, and roughness as measured by International Roughness Index (IRI).

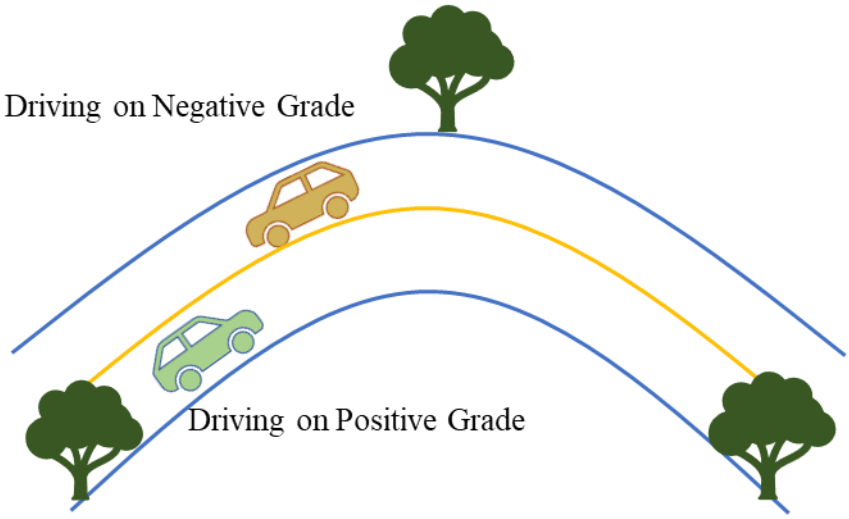
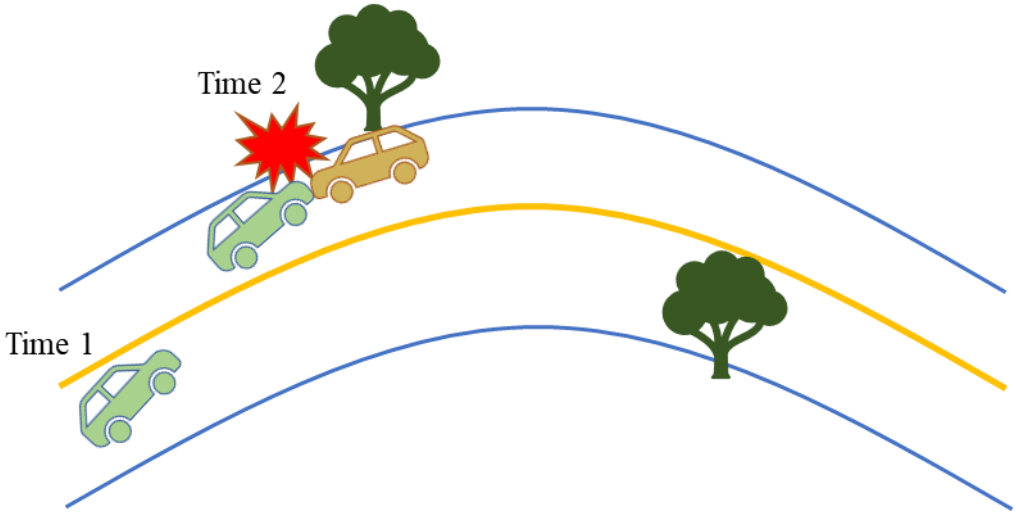


Figure 1 A sketch illustrating the sign of grade for each direction of travel (longitudinal profile)

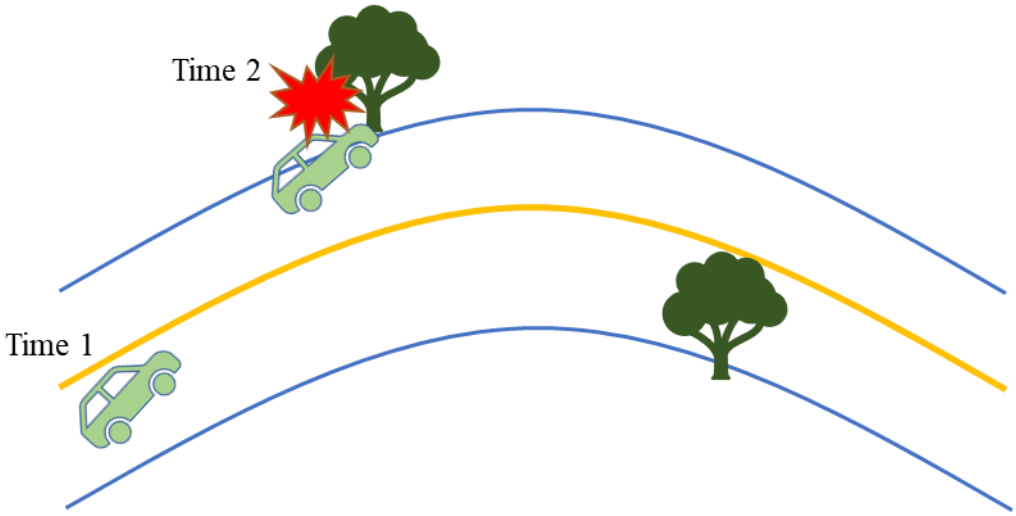
In addition, for TLTW roads, a separate analysis by direction allows the assignment of a head-on crash to the direction of travel of the vehicle that according to the police report caused the crash. This helps in the development of more realistic and accurate models, as the crash is more likely to have been caused by attributes in that direction of travel. Figure 2 highlights the importance of directional analysis on the assignment of crashes to segments. In both cases, the figure shows a roadway departure to the left side but in part a) the departure results in a head-on crash, while in part b) the car crosses the opposite lane and ends up colliding with a fixed object simply because no car was passing in the opposite direction. In this study, both crashes are assigned to the direction of travel of the vehicle causing the crash.

Using recently collected roadway characteristics (e.g., traffic, geometry, etc.) in each direction, this study will explore the importance of considering the direction of travel

in modeling crash frequencies of some roadway departures that lend themselves to directional analysis.



a)



b)

Figure 2 A sketch illustrating different types of RwD crashes to the left (plan view) a) A head-on collision and b) A fix object collision

1.2.3 The general geometric environment of the analysis segment

To the best of the writer's knowledge, crash frequency models have been developed based only on the attributes of the segment on which the crash occurred. This means that for each segment, only that segment's attributes are considered as independent variables to predict the total number of crashes. However, intuitively, it may be expected that the general geometric environment of the roadway portion prior to where the segment is located could affect the frequency of crashes. As recognized in highway design practices (design consistency), the safety of the road may be affected not only by the mean and/or variance of a given attribute (e.g., curvature) within the segment but also by changes relative to the value on a longer portion of the road where the section is located. In fact, design consistency practices recognize this by trying to vary features gradually (equivalently, reducing the variability of some roadway features such as curvature). For instance, two curves with identical curvatures and lengths can have significantly different effects on RwD crash frequency if one is preceded by a straight roadway while the other is preceded by a winding alignment (Figure 3). Notice that consideration of the roadway environment prior to the study segment necessarily requires a directional analysis as discussed in the previous section.

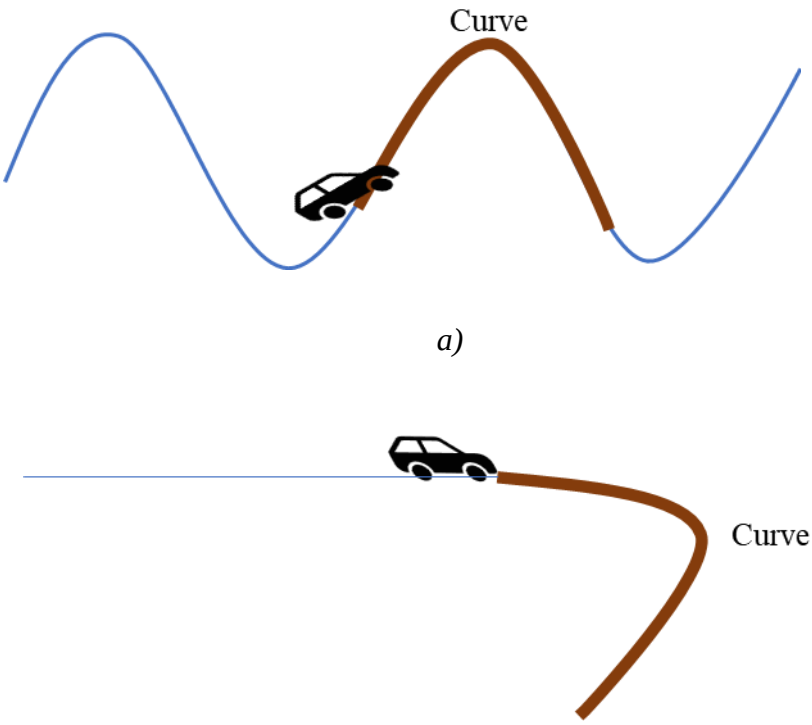


Figure 3- A sketch illustrating the role of the general geometric environment

a) curve on a winding road and b) curve on a straight road (plan view)

The consideration of the mean and standard deviation of some features of the general geometric environment, such as curvature, relative to the mean and standard deviation of the same feature in the segment may be proxies for higher speed of vehicles on straight segments before the curve or longer drivers' reaction time (drivers may not expect to face a curvy segment on a generally straight road). In this study, it is proposed to evaluate the effects of the geometric environment of a longer portion of the roadway on which the study section is located.

1.3 Research objectives

This research presents the development of crash frequency models with roadway related explanatory variables for TLTW state roads in Hawaii, and it identifies contributing factors to the frequency of RwD crashes. The primary objective of the dissertation is to explore the sensitivity of the estimated models to the factors discussed in the previous section, namely, a) the effects of segment length, b) the effects of roadway direction of travel, and c) the effects of features of the general geometric environment of a longer portion of the road where the study segment is located.

1.4 Dissertation outline

The next chapter of this dissertation, Chapter 2, presents a brief review of the literature on crash frequency modeling and the RwD crashes. Chapter 3 describes the different data sources followed by the steps required to prepare the data for modeling. It also presents the methodologies used to evaluate the effect of the data synthesis on crash frequency models. Lastly, it provides a detailed description of the explanatory variables used for modeling. Chapter 4 presents the model estimation results including their interpretations, evaluations, and comparisons. Finally, Chapter 5 presents the study conclusions and recommendations.

CHAPTER 2: BACKGROUND

2.1 Crash frequency models

The total number of crashes observed on roadway segments is an example of count data, with only non-negative integer outcomes. A standard regression model is inadequate for modeling count data since it fails to limit its predictions to non-negative integers (i.e., predicting values that are non-integer and/or negative). Count data is properly modeled by Generalized Linear Models (GLM), such as Poisson and negative binomial regression models since they are suitable to predict the probability of observing rare events (e.g., a certain number of RwD crashes) very well. Some of the properties of GLMs are *(8)*:

- They do not assume a linear relationship between the dependent variable and the independent variables.
- The dependent variable does not need to be normally distributed.
- Errors need to be independent but not normally distributed.
- Homoscedasticity is no longer assumed.

Traffic crashes are complex phenomena since they are the results of interactions between humans, roadway characteristics, vehicles, and environmental conditions. There has been steady progress in developing new mathematical approaches to capture the complexity of crashes for many years *(9)*. Different methodologies have been applied in crash frequency analyses such as Poisson*(10)*, negative binomial *(11)*, zero-inflated negative binomial *(12)*, Conway-Maxwell Poisson *(13)*, negative multinomial *(14)*, and negative binomial Lindley model *(15)*. Roadway characteristics may have heterogeneous

effects across the observations due to the unobserved time-varying environmental conditions as well as the heterogeneous reaction of drivers to the roadway features *(15).* Since it is not possible to collect all the required data to predict the likelihood of crashes on roads, various statistical approaches such as random parameter negative binomial *(16)* or mixed-effects negative binomial models *(17)* are introduced to consider the unobserved heterogeneity in crash data analysis *(18).*

Lord and Mannering reviewed different methodological alternatives for modeling crash frequencies *(19)*, including Poisson, Negative Binomial, and Zero-Inflated models (among others). These, together with mixed-effects models, are briefly summarized below.

2.1.1 Poisson regression

Poisson regression has been used as an alternative to model the total number of RwD crashes on roadway segments. In a Poisson regression model, $P(n_i)$ the probability of *n* crashes occurring on segment i is given by:

$$P(n_i) = \frac{exp(-\lambda_i)\lambda_i^{n_i}}{n_i!} \quad (1)$$

$$\lambda_i = exp(\boldsymbol{\beta}\boldsymbol{x_i}) \quad (2)$$

$P(n_i)$: probability of n crashes occurring on segment i

λ_i: expected crash frequency for segment i

$\boldsymbol{x}_i$: vector of explanatory variables for segment i (the i^{th} row of the design matrix $\boldsymbol{X}$)

$\boldsymbol{\beta}$: vector of estimable coefficients

2.1.2 Negative binomial regression

A significant limitation of Poisson regression is related to its use of the Poisson distribution, which requires that the mean of the count data be equal to its variance. If the variance is larger than the mean, the data is over-dispersed. For such data, a negative binomial model is more appropriate *(20)*. As in Poisson regression, in negative binomial regression, the expected crash frequency for segment i is modeled as a function of explanatory variables collected in a vector $\boldsymbol{x}_i$ such that:

$$\lambda_i = exp(\boldsymbol{\beta}\boldsymbol{x}_i + \varepsilon_i) \tag{3}$$

where ε_i is the error term for segment i. It is commonly assumed that $\exp(\varepsilon_i)$ is Gamma distributed with a mean of one and a variance of α. The addition of the error term ε_i allows the variance of n_i to differ from the mean:

$$VAR[n_i] = E[n_i]\left[1 + \alpha E[n_i]\right] \tag{4}$$

The negative binomial distribution for predicting the probability of observing a crash frequency n_i is:

$$P(n_i) = \left(\frac{\Gamma\left(n_i + \frac{1}{\alpha}\right)}{n_i!}\right) \times \left(\frac{\frac{1}{\alpha}}{\frac{1}{\alpha} + \lambda_i}\right)^{\frac{1}{\alpha}} \times \left(\frac{\lambda_i}{\frac{1}{\alpha} + \lambda_i}\right)^{n_i} \tag{5}$$

Where $\Gamma(.)$ is the Gamma function and α is the overdispersion parameter. The vector of parameters for this model ($\boldsymbol{\beta}$) can be estimated by maximum likelihood. Winkelmann

explains that ε_i, the error term in negative binomial regression, can be considered as a random-effect term to model the individual heterogeneity *(21)*.

2.1.3 Zero-inflated negative binomial regression

The zero-inflated negative binomial model addresses the problem of having a high fraction of zeros in the response variable. Observation of zero events during the observation period can result from two separate representations, one set of observations that are necessarily zeroes and another set with a small probability of observing zeros. In terms of crash data analysis *(20)*, these representations imply that either the segment is absolutely safe and the probability of crash occurrence is zero, or that the failure to observe any crashes in the period of study comes from the representation with a low probability of observing a zero crash. Following the notation by Winkelman *(21)*, zero-inflated models combine a binary variable C_i with a standard count variable n_i^* that is given by:

$$n_i = \begin{cases} 0 & if\ C_i = 1 \\ n_i^* & if\ C_i = 0 \end{cases} \quad (6)$$

If the probability that $C_i = 1$ (stating the segment is absolutely safe) is denoted by w_i, the probability function of n_i is:

$$P(n_i) = w_i d_i + (1 - w_i) g(n_i) \quad (7)$$

where $d_i = 1 - \min\{n_i, 1\}$ and $g(n_i)$ is a count data probability function like Poisson or a negative binomial probability function *(21)*. Usually, a binary logit model is used to model the binary status of the two parts of a zero-inflated model. Therefore, the zero-

inflated negative binomial probability density function can be expressed as a combination of the two processes as below:

$$P(n_i) = w_i + (1 - w_i)\left(\frac{\frac{1}{\alpha}}{\frac{1}{\alpha} + \lambda_i}\right)^{\frac{1}{\alpha}} \qquad if \quad n_i = 0 \qquad (8)$$

$$P(n_i) = (1 - w_i) * \left(\frac{\Gamma\left(n_i + \frac{1}{\alpha}\right)\left(\frac{\frac{1}{\alpha}}{\frac{1}{\alpha} + \lambda_i}\right)^{\frac{1}{\alpha}}\left(1 - \frac{\frac{1}{\alpha}}{\frac{1}{\alpha} + \lambda_i}\right)^{n}}{\Gamma\left(\frac{1}{\alpha}\right) n_i!}\right) \quad if\ n_i = 1,2,3,\ldots \qquad (9)$$

Lord et al. provide further notes on the detailed application of this model for safety data analysis *(22) (23)*.

2.1.4 Generalized linear mixed-effects model- negative binomial

The negative binomial formulation considers a constant estimated coefficient for each explanatory variable across all the observations. However, it is widely accepted that there are many threads of unobserved heterogeneity in crash data analysis that may change the effect of explanatory variables for different observations *(18)*. Therefore, to account for the unobserved heterogeneity in crash data analysis, it is proposed to add a random term (i.e., a random effect) to the parameter of each explanatory variable l as follows:

$$\beta_{il} = \beta_l + \varphi_{il} \qquad (10)$$

where β_{il} is the parameter on the l^{th} explanatory variable for observation i, β_l is the mean parameter on the l^{th} explanatory variable across all the observations (i.e., fixed effects), φ_{il}

is a randomly distributed scalar term that captures unobserved heterogeneity across the observations (i.e., random effects), and its distribution can be assumed by the analyst (e.g., normal, lognormal or uniform distribution). Basically, if the variance of the chosen distribution is not significantly different from zero, only the fixed parameter would remain in the model. Therefore, the model embraces both the fixed effects and the random effects, and the likelihood function can be written as:

$$L = \prod_{\forall i} \int_{\varphi_i} P(n_i|\varphi_i)\, g(\varphi_i) d\varphi_i \quad (11)$$

where the $g(\varphi_i)$ is the probability density function of φ_i. Since the required numerical calculation of this likelihood function is cumbersome, a simulated likelihood approach *(24)* (using Halton sequences to limit the required number of draws) is utilized to estimate the parameters in the literature *(16) (25) (26)*.

The above approach is called random parameter negative binomial regression model in the crash frequency modeling literature *(16) (27) (26)*, which is equivalent to what is known as Generalized Linear Mixed Effects Model-Negative binomial (GLMM-NB) in the terminology commonly used in statistics *(28) (17) (29)*. Generally, a mixed-model contains both fixed effects and random effects. Fixed effects are parameters which can be associated to the entire population explaining the behavior of the population means, but random effects are associated with individual sections which are drawn at random from an entire population *(28)*. GLMM-NB can be a suitable extension to model RwD crashes. The GLMM-NB develops a general model by including fix-effects describing the average variation of the mean number of crashes on segment *i* with a set of covariates gathered in

a design matrix $\boldsymbol{X}$, as well as random-effects related to explanatory variables for each segment collected on a matrix $\boldsymbol{Z}$ used to consider the unobserved heterogeneity effects across the observation. A generalized linear mixed effects model (using crash data analysis terminology) is:

$$\lambda_i = \boldsymbol{x}_i \times \boldsymbol{\beta} + \boldsymbol{z}_i \times \boldsymbol{\varphi_i} + \varepsilon_i \quad (12)$$

$\boldsymbol{x_i} \times \boldsymbol{\beta}$ is the fixed effects term, and $\boldsymbol{z_i} \times \boldsymbol{\varphi}_i$ is the random effects term,

λ_i is the mean number of crashes for section i,

$\boldsymbol{x_i}$ is a $(1 \times p)$ row vector consisting of the i^{th} row of the design matrix $\boldsymbol{X}$ collecting the information of the values of the explanatory variables for section i,

$\boldsymbol{\beta}$ is a $(p \times 1)$ column vector of the fixed-effects regression coefficients,

$\mathbf{z}_i$ is a $(1 \times q)$ row vector collecting the information related to random effects for section i (the i^{th} row of matrix $\boldsymbol{Z}$),

$\boldsymbol{\varphi}_i$ is a $(q \times 1)$ vector of random effects for segment i,

ε_i is the error term for section i,

p is the number of explanatory variables in $\boldsymbol{X}$

q is the number of explanatory variables in $\mathbf{Z}$.

Similarly, the mixed-effects model (Equation 12), can be extended with grouped data to model the response variable as a function of explanatory variables by considering the correlation between observations in each group and the variation between groups that might affect the response variables *(28)*. This separation of the two sources of variation results in consistent and efficient variance standard errors and in turn, in more reliable statistical results for the parameter estimates. In the context of crash frequency modeling,

researchers cannot control what constitutes a group. Nevertheless, it may be possible to group sections into sets with similar values of explanatory variables.

To sum up, although there has been a steady advancement in crash frequency modeling, the ability of these models is limited by the available databases used to estimate their parameters *(19)*. Moreover, no general rule establishes the superiority of one methodological approach over another for crash data analysis *(9)*.

2.2 Rate models

The total number of observed RwD crashes on a roadway segment depends on the length of the segment, AADT, and the study period (i.e., the total number of years). The dependent variable depends on the size of variables that determine the number of opportunities for the event (i.e., RwD crashes) to occur *(17)*. Following the notation in Faraway*(17)*, it is possible to relate the Poisson model with a log link back to a linear model for the ratio response. Therefore, Equation 2 can be rearranged for segment i as follows:

$$\log\left(\frac{\lambda_i}{Number\ of\ years_i \times Length_i \times AADT_i}\right) = \boldsymbol{\beta x}_i \quad (13)$$

$$log(\lambda_i) = log(Number\ of\ years_i) + log(Length_i) + log(AADT_i) + \boldsymbol{\beta x}_i \quad (14)$$

Therefore, the general formulation of a rate model for Poisson regression is:

$$\lambda_i = e^{\beta_0 + \beta_1 (Variable\ 1) + \beta_2 (Variable\ 2) + \ldots + \beta_n (Variable\ N)} \quad (15)$$

where,

$$\beta_0 = Intecept + log(Length) + log(Years) + log(AADT) \quad (16)$$

Rate models are easily implementable in R *(30)*, the software used for model estimation in this study, with the help of the offset command. This command forces the model to assume a fixed coefficient equal to one for variables specified in offset commands (as required by Equation 16 for the logarithms of the three variables $Length$, $Years$, and $AADT$). Modeling the frequency of RwD crashes as a rate model enables the comparison of estimated parameters of separately developed models with different segment length. It also facilitates the comparison of the effect of explanatory variables on the safety of the roads with widely different traffic levels. Furthermore, it permits the estimation of models with a different number of years of data available for different segments. This is not the case in the database used in this study since the same number of years is available for each segment, but it would be if the data is further segmented based on some safety treatments such as the use of centerline rumble strips. For many segments, the ten years of available data would need to be split in a before and after treatment; thus, the same segment would appear twice in the dataset, once without the treatment and once with the treatment with fewer than ten years for each.

It is important to emphasize that although the model estimates the parameters using the total number of RwD crashes, the parameters correspond to the crash frequency rate model.

2.3 RwD crashes

As mentioned earlier, the synthesis of the data is an essential step in developing crash frequency models. In fact, it is hypothesized that it may have a substantial effect on the results and lead to poor inferences if careful attention is not paid to its details. For instance, a review of the roadway segmentation used in previous studies indicates that the two conventional approaches to split the roads into smaller analysis segments are used without justifying the effect on the results. For instance, Peng et al. investigated the relationship between single vehicle run off the road crashes and the geometric characteristics of rural two-lane roads by using a negative binomial model *(31)*. In that study, only five independent variables were used to model the frequency of run off the road crashes. The author did not explain how the 245.3 miles of roads are divided into 501 roadway segments, while the segment length varied from 0.1 to 11.1 miles. Anastasopoulos et al. *(16)* studied factors affecting rural interstate crashes in Indiana by defining 322 homogenous segments based on shoulder characteristics, pavement type, median characteristics, number of lanes, and speed limit. The segment length in this study varied from 0.1 to 11.53 miles. The authors did not justify why only those variables were selected, and how the homogeneity of a long segment could be warranted if only those attributes remained constant (e.g., what if the curvature or grade changes considerably along a long segment).

Similar issues are observed in studies that employed the fixed length approach. Lee and Mannering *(12)* employed zero-inflated negative binomial to study the frequency of run-off the road crashes. They used 120 segments of equal 0.5-mile length over a 60 miles

road in the state of Washington. They did not explain why they chose a 0.5-mile segment length. In general, no information was found in the literature investigating the effect of segment length, as a part of the data preparation process, on the estimation of the RwD crash frequency models and their contributing factors. Similarly, no information was found in the literature considering the effect of the direction of travel on TLTW roads, or the general geometric environment of the analysis segment on the frequency of RwD crashes.

Generally, few research studies on crashes have been conducted in the State of Hawaii so far. They focused mostly on the role of land use and spatial distribution of sociodemographic characteristics on all types of crashes rather than developing a statistical model to predict crashes based on the roadway characteristics (e.g., traffic, geometry, and inventory databases). For instance, Kim et al. investigated the interplay between demographic, land use, roadway accessibility variables and types of crashes in Oahu *(32)*. Meanwhile, other methodologies have been employed for studying RwD crashes in Hawaii. Hashemi and Archilla *(33)* analyzed RwD crashes using Bayesian analysis to investigate the most prevalent versus the most probable circumstances of RwD crashes. Also, they explored the application of the Classification an Regression Trees (CART) as an exploratory analysis tool to identify the RwD crashes' contributing factors on the Island of Oahu in Hawaii *(34)*. Hence, these studies are limited in a sense that none of them opted to reveal the quantitative relationship between the frequency of crashes and their contributing factors based on a robust statistical model that provides enough insights for decision-makers to act.

Therefore, the primary intent of this dissertation is to provide insight into the factors that affect the frequency of RwD crashes on TLTW state roads in Hawaii. To do so, the effect of segment length, the directional analysis and the general geometric environment of the analysis segment are explored to fill the gap in the literature. The results of this analysis highlight the necessity of careful data synthesis process before developing a model. Also, different statistical methodologies are explored and compared in terms of their statistics (e.g., log-likelihood, AIC, and BIC), the goodness of fit (as measured by an observed vs. predicted graphs), and other practical considerations. As discussed later, all models provide relatively good representations of the data which makes the tradeoff between model assumptions and relative ease of use difficult.

The next section of this research fully explains the methodology. It describes the data sources for this study followed by the key steps required for the data preparation and the data synthesis process.

CHAPTER 3: DATA DESCRIPTION & METHODOLOGY

This dissertation investigates the frequency of RwD crashes using ten years of crash data from the State of Hawaii, together with detailed roadway characteristics (e.g., traffic, geometry, etc.). This section describes the different data sources and explains the required steps for data preparation. Then, it explains the approaches used to process the data to accomplish the objectives of this dissertation. Finally, it provides some descriptions of explanatory variables.

3.1 Data sources

This research uses data from different sources to conduct the analysis. A brief explanation of each source is provided in the following sections.

3.1.1 Motor vehicle accident reports

The State of Hawaii motor vehicle accident reports for ten consecutive years (2005-2014) were obtained from HDOT's highway division traffic branch. This database was used to extract the RwD crashes on the islands of Hawaii (Hawaii, Oahu, Maui, and Kauai).

The motor vehicle accident reports include information at three levels: crash level, vehicle level, and passenger level. In contrast to crash severity models[1] that can use the

[1] Crash severity models predict the probability that corresponds to each level of severity based on the attributes of driver, passengers, vehicles, and the other general circumstances of crashes that are available in police crash reports.

above information at a disaggregate level (i.e., for each crash), crash frequency models cannot fully use such information since the unit of analysis is the roadway segment.

3.1.2 Roadway characteristics

Roadway characteristics were obtained from HDOT's highway division planning branch. Different features are stored in separate datasets by HDOT. So, an important challenge was to combine the data from different datasets into a single comprehensive dataset for analysis and then to find the proper value for each segment to study the frequency of RwD crashes. The available data were reviewed several times to ensure all explanatory variables (e.g., AADT, curvature, grade, lane width, lane type, IRI, and shoulder width) that may reasonably affect the frequency of the RwD crashes are included in the model to reduce the unobserved heterogeneity *(18)*.

3.1.3 GIS data

The GIS shapefile of the state roads was obtained from HDOT's highway division planning branch. In addition, other shapefiles such as coastline were obtained from the State of Hawaii Office of Planning website *(35)*.

3.1.4 Video log data

The video log data of the TLTW state roads on Hawaii, Oahu, Maui, Kauai, and (from existing HDOT records) were used to clarify any unclear point in the available databases.

3.2 Data preparation

3.2.1 Roads selection

The selected roads are all the TLTW state roads in the State of Hawaii. The TLTW roads were extracted based on a dataset that includes the state road characteristics such as the total number of lanes in each direction. In addition, the GIS shapefile of the state roads was used to identify their geographic locations within the islands. The portion of the state highway network selected for the study covers more than 600 centerline miles (more than 1200 lane-miles), which is slightly less than half of the state highway network. The selected state roads include approximately 325.8 centerline miles in the island of Hawaii (state roads 11, 19, 130, 190, 200, 240, 250, and 270), 76.2 centerline miles in Oahu (state roads 72, 83, 93, 99, 750, and 930), 115.6 centerline miles in Maui (state roads 30, 36, 37, 360, 377, and 378), and 82.4 centerline miles in Kauai (state roads 50, 56, 550, 560, and 580).

Table 1 presents more details of the selected roads including the road number, the beginning mile point (BMP), the ending mile point (EMP), and the name of the island. It worth mentioning that on roads 19 and 30 there are segments in between the selected segments that are not TLTW roads. Therefore, those segments are excluded from the analysis.

Figures 4 through 7 show GIS maps of the selected TLTW state roads on each island. The state roads are depicted with their road number. These figures demonstrate that a large portion of TLTW state roads includes coastline roads.

Table 1- TLTW state roads in the State of Hawaii

Roads Number	BMP	EMP	Length	Island
11	8.39	109	100.61	Hawaii
19	8.79	52	43.21	Hawaii
19	58	86.51	28.51	Hawaii
130	4.01	21.63	17.62	Hawaii
190	0.1	36.88	36.78	Hawaii
200	0	43.22	43.22	Hawaii
240	0	9.59	9.59	Hawaii
250	0	19.27	19.27	Hawaii
270	0	27	27	Hawaii
72	2.08	13.23	11.15	Oahu
83	0	39.5	39.5	Oahu
93	12.71	19.52	6.81	Oahu
99	0	6.52	6.52	Oahu
750	0.84	7.21	6.37	Oahu
930	0	5.93	5.93	Oahu
30	0	4.59	4.59	Maui
30	6.68	19.58	12.9	Maui
30	26.09	41.61	15.52	Maui
36	3.301	16.21	12.909	Maui
360	0	34.82	34.82	Maui
37	5.63	21.34	15.71	Maui
377	0	9.13	9.13	Maui
378	0	10.09	10.09	Maui
50	1.974	32.91	30.936	Kauai
56	7.384	28.11	20.726	Kauai
550	0	14.07	14.07	Kauai
560	0	10	10	Kauai
580	0	6.68	6.68	Kauai

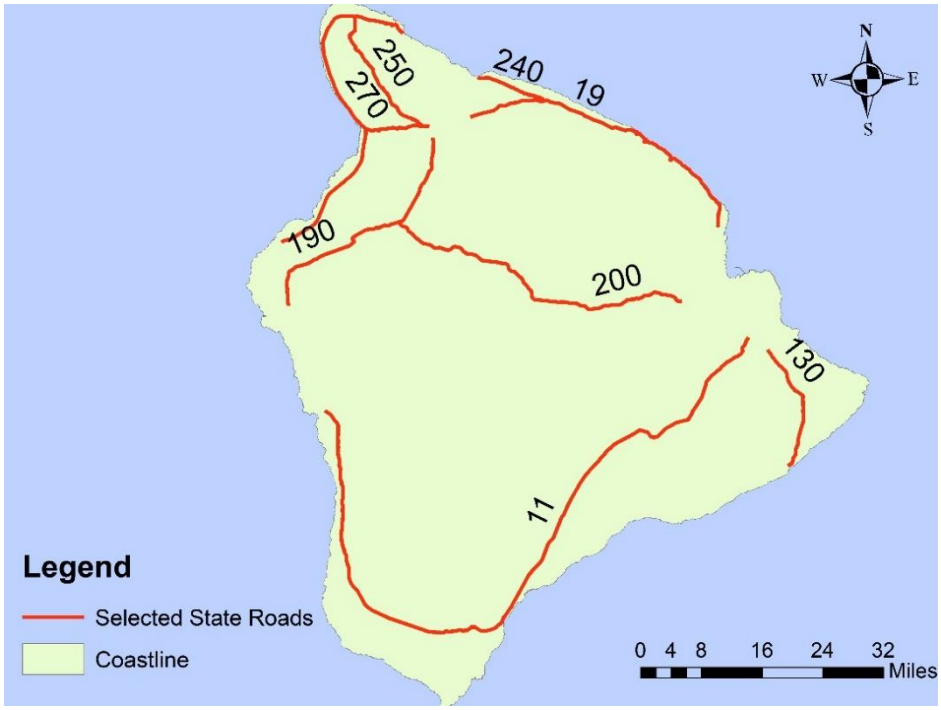


Figure 4 Two-lane two-way state roads on the island of Hawaii

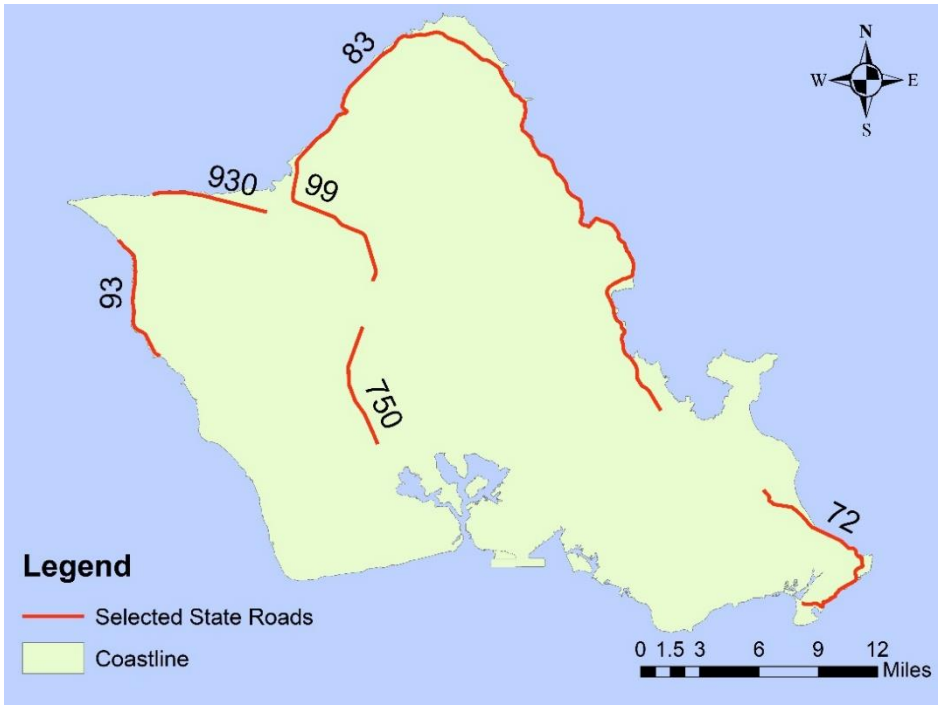


Figure 5 Two-lane two-way state roads on Oahu

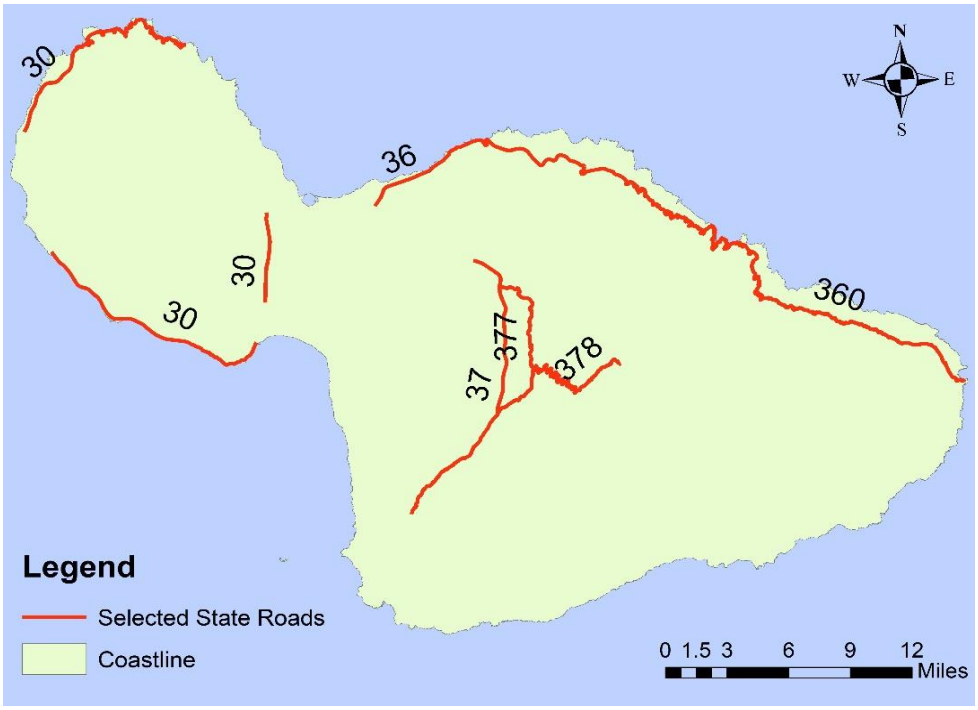


Figure 6 Two-lane two-way state roads on Maui

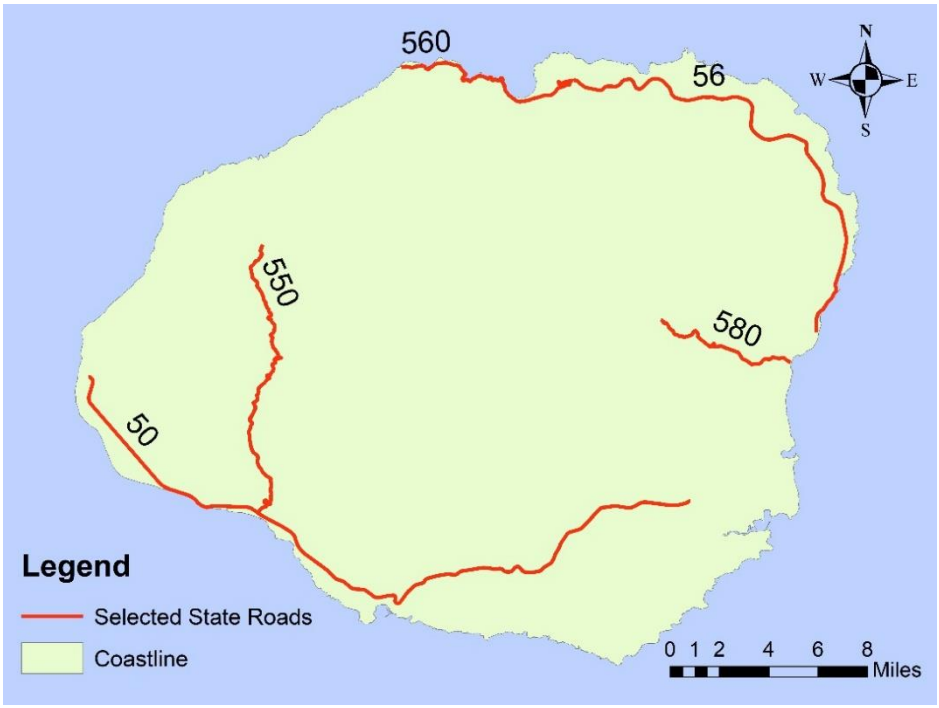


Figure 7 Two-lane two-way state roads on Kauai

3.2.2 Extracting the RwD crashes

The RwD crashes were extracted using the FHWA definition: a non-intersection crash in which a vehicle crosses an edge line, a centerline, or otherwise leaves the traveled way *(1)*. After a detailed examination of the motor vehicle crash reports (Appendix 6.1 presents the State of Hawaii motor vehicle accident report form), a list of first harmful events identifying whether a crash is a RwD was selected. This task was completed with guidance from HDOT's highway division traffic branch personnel. Table 2 presents a list of first harmful events that are considered as resulting in RwD crashes. It also shows the first harmful events differentiating whether a crash is a RwD to the right side or to the left side of the road. Afterward, to maintain the compatibility with the FHWA RwD crashes definition, the locations of all crashes are filtered to the non-intersection crashes, and the first harmful event of crashes are filtered to the list provided in Table 2.

3.2.3 Crash geolocation

Since the crash data for years 2005 to 2011 were not geolocated, a substantial initial effort of this study was spent in the geolocation of crashes (i.e., identifying the location of a crash with a unique latitude and longitude). These locations were then converted to the actual mile point on the roads, and then the direction of travel for all the crashes (2005-2014) was identified manually before proceeding to statistical data analysis. These time-consuming processes were completed with software such as Google Earth, Google Maps and Arc-GIS. Table 3 presents an example of the location information reported by the police in the crash report.

Table 2 A list of first harmful events resulting in RwD crashes

Type of Collision	RwD to the right-side	RwD to the left-side
Non-Collision	02 Overturn/Rollover Off Roadway 03 Submersion 06 Ran Off Roadway	11 Cross Median/Centerline
Collision with Object/Animal	21 Guardrail Face, 22 Guardrail End, 23 Culvert, 24 Ditch, 26 Bridge Pier or Support, 27 Bridge Rail, 28 Building, 29 Tunnel, 30 Curb, 31 Embankment /Retaining Wall, 32 Fence, 33 Utility Pole/Light Support, 34 Traffic Signal/Sign Post, 35 Other Post/Pole/Support, 36 Impact Attenuator/Crash Cushion, 37 Concrete Traffic Barrier, 38 Other Traffic Barrier, 39 Tree (Standing), 40 Hydrant, 41 Mailbox	
Collision with Bicycle or Moped	71 Riding in Bikeway 74 Riding off Roadway Direction	-
Collision with MV in Transport (Except Moped)	-	80 Head-On, 83 Sideswipes Opposite, 85 Angle Opposite Direction
Collision with MV - Other	102 Parked MV	-

Table 3 An example of a crash location in police crash report

Name of the Road	Road No.
KAMEHAMEHA HWY	83
Distance and Direction	**Refer Road (Intersection, Etc.)**
213 ft., North	WAIAHOLE VALLEY RD

RwD crashes were geolocated following these steps:

1. Finding the intersection of the “Name of the Road” and the “Refer Road” (reference road), fields in Table 3, on the map.
2. Finding the crash location on the main road using the specified distance and direction provided with respect to the “reference road.”

3. Using the direction before the crash of the first car in the report (a number from 1 to 8 representing N, NE, E, SE, S, SW, W, and NW) to determine the direction in which the crash occurred.
4. Double-checking the result with the diagram of a crash drew by a police officer for some unclear cases.

For the geolocation process, the WGS 84 geodetic datum was used to show the specific latitude and longitude for the crashes. These locations were imported into Arc-GIS. To properly show the location of the crashes relative to the road network represented in a shapefile of state roads acquired from HDOT, the data were projected into the CS North American 1983 HARN system. Finally, the actual GIS locations of the crashes were converted to the road mile point (distance from the zero-mile point along the alignment).

3.3 Data synthesis process

This research aims to develop statistical models that predict RwD crashes on roadway segments. Therefore, the total number of RwD crashes on each segment is the dependent variable, and roadway characteristics (e.g., traffic, roadway geometrics) are the independent variables. The independent variables include both continuous and indicator variables (i.e., dummy variables). Hence, all the collected geometry and inventory attributes should be assigned based on the unit of analysis, which is the segment.

A total number of 4,604 RwD crashes were observed on TLTW state road during the study period (i.e., 2005-2014) in Hawaii[1]. Figure 8 presents the distribution of RwD crashes based on their severity according to Hawaii's police crash reports. The severity of each crash is assigned based on the most severe injury in each crash. The possible severity levels are fatal, incapacitating injury, non-incapacitating injury, possible injury[2], and none[3]. Figure 8 shows that the total number of RwD crashes with severe injuries (fatal and incapacitating injuries) are very limited. Therefore, all the RwD crashes without regards for the severity are used to develop the crash frequency models. Meanwhile, developing a comprehensive dataset including the total number of RwD crashes and the roadway characteristics was a challenge, especially since the roadway characteristics data were collected separately with a different number of segments, lengths, beginning and ending mile points.

[1] The state of Hawaii has 1.42 million population (based on census estimates for 2018, Source: www.wikipedia.org), and it ranked 41th between 50 states in US.

[2] According to FHWA, "A possible injury is any injury reported or claimed which is not a fatal injury, incapacitating injury or non-incapacitating injury. Inclusions: Momentary unconsciousness, claim of injuries not evident, limping, complaint of pain, nausea, hysteria" *(43)*.

[3] Equivalent to the Property Damage Only (PDO) crashes or non-injury crashes.

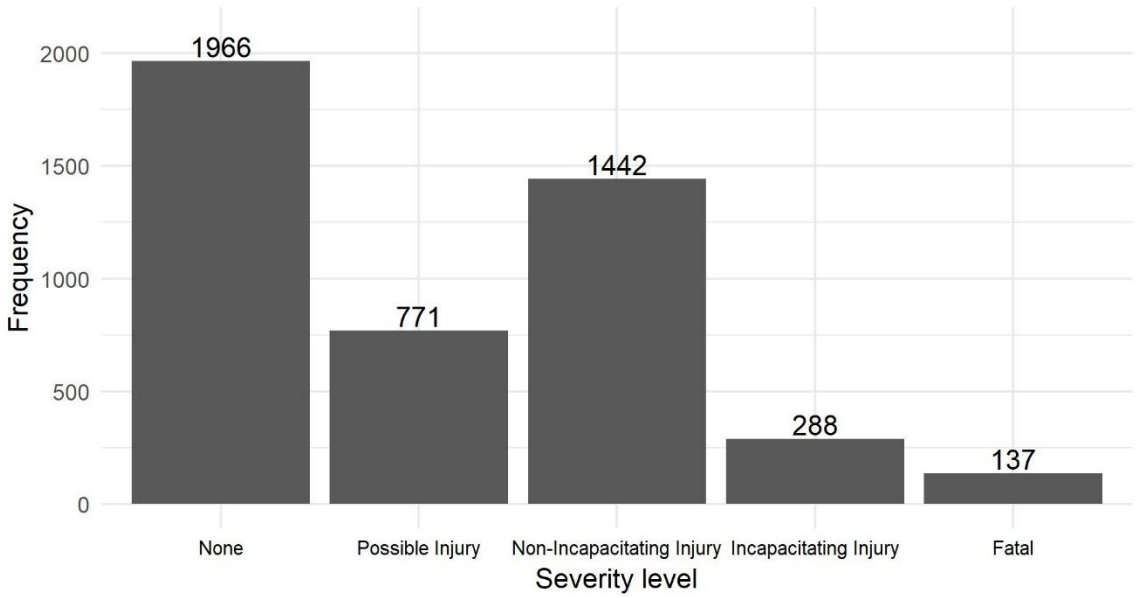


Figure 8 A histogram illustrating the distribution of RwD crashes by their severity

3.3.1 Segment length

In crash frequency modeling, each segment corresponds to a data point that is used to estimate the parameters of a generalized linear regression model. Therefore, each segment should be as homogenous as possible for two primary reasons. First, the locations of crashes are assumed anywhere along the length of the segment. Second, crashes are predicted based on a set of attributes whose values are considered representative for each segment. Hence, the analysis is under the influence of roadway segmentation that is defined by the analyst. Selecting a specific type of road may contribute to maintaining the homogeneity of segments. In this research, this already has been done by limiting the selected roads to TLTW state roads in the State of Hawaii. However, shorter segments tend to be more homogeneous than longer segments; thus, one way to maintain the homogeneity

is to keep the segment's length short. Here some guidance is provided to keep the segments relatively homogeneous with the selection of an adequate segment length.

It is believed that selecting an appropriate segment length is important because if the length is too short, then most segments would have zero or just one crash, thus defeating the purpose of crash frequency modeling. Also, with segments that are too short, it may be difficult to capture the effects of some contributing factors unless general road environment characteristics are explicitly included in the dataset. An excellent example of this is a crash caused by a driver who started to lose control of the vehicle on a sharp curve but who crashed on an adjacent tangent. In this study, it is hoped to solve this issue either with the selection of the segment length and/or with the consideration of the geometric environment of the road where the section is located.

On the other hand, it is also envisioned that a limit for segment length is needed to reduce the information loss caused by aggregating the data over an extended length of road (e.g., a segment with a sharp curve in a middle of a long straight segment would have a small average curvature, so the effect of curvature may not be captured accurately).

Crash frequency histograms derived from different segment lengths expose the effect of segment length on the distribution of crashes. Generally, longer segment lengths yield relatively flat histograms while shorter segment lengths generate histograms mostly with zero and one frequencies. Starting from shorter to longer segment lengths, the mode of histograms changes from zero to larger values. The percentage of segments with a specific number of crashes (e.g., zero, one, two) can also be obtained from the histograms.

Generally, the histograms become flatter as the number of segments with zeros crashes decreases and the number of segments with a higher number of crashes increases.

On the other hand, the goal is to maintain the homogeneity of the segment by keeping them short. In this regard, it might be possible to suggest a maximum segment length that is commensurate with the type of roads and their attributes including the geometric features. The rationale is that shorter segments would result in too many segments with no crashes (too many zeros), while excessively averaging of roadway characteristics would be obtained with longer segment lengths.

Therefore, separate crash frequency models are estimated using different fixed-length (e.g., 0.1-, 0.2-, 0.3-, 0.5-, 1.0-, and 2.0-mile segment lengths) to evaluate the effect of segment length on the estimation of crash frequency models.

For each segment length, a new set of segments was generated before proceeding to the modeling. For example, for a 0.1-mile segment length, a 100-mile road would be divided into 1000 segments. For a 0.2-mile segment length, the same road would result in 500 segments. For each segment length, the total number of crashes on each segment was calculated. Necessarily, the summation of the total number of crashes in both cases (i.e., the summation of 1000 values versus 500 values) should be equal. Table 4 presents an example of this calculation.

Table 4 An example of different roadway segmentations

a) 0.1-mile segmentation

Segment number	BMP	EMP	Crashes
1	0	0.1	1
2	0.1	0.2	2
3	0.2	0.3	0
4	0.3	.4	1
⋮	⋮	⋮	⋮
997	99.6	99.7	0
998	99.7	99.8	2
999	99.8	99.9	3
1000	99.9	100	2
Total number of Crashes			**150**

b) 0.2-mile segmentation

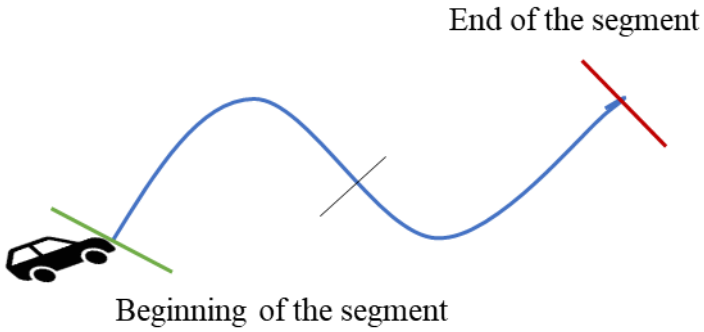
Segment number	BMP	EMP	Crashes
1	0	0.2	3
2	0.2	0.4	1
⋮	⋮	⋮	⋮
499	99.6	99.8	2
500	99.8	100	5
Total number of Crashes			**150**

Considering a reasonable relationship between the dependent variable (frequency of the RwD crashes) and independent variables (e.g., geometric features) is of most interest in this study since it is believed it can affect the explanatory power of some variables in the estimated models. The independent variables, however, should be compatible with the unit of analysis. Therefore, for each segment length, weighted averages are used to calculate the values of explanatory variables to capture the effect of the segment length on the frequency of RwD crashes.

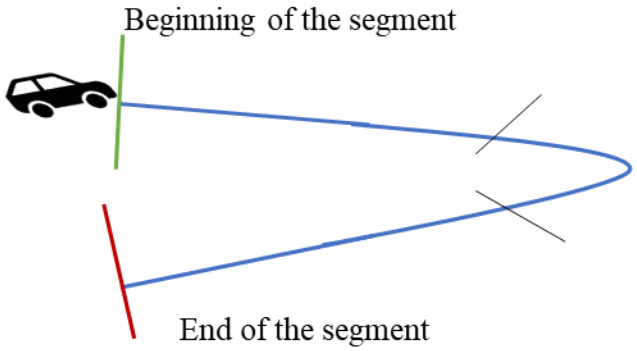
Moreover, for some independent variables such as curvature, not only its mean but also its standard deviation is considered. The reason is illustrated in Figure 9, which shows two segments with the same mean of curvature, but with different standard deviations. The first segment is composed of several gentler curves whereas the latter is composed of basically two straight segments with a sharp curve in between. With all else equal, it is still reasonable to expect different crash frequencies in the two segments even though the average curvature is the same. Therefore, a weighted standard deviation is calculated to capture this feature.

Although variables such as curvature and grade are continuous, on a given segment they have only a finite number of values (and even when that is not the case, such as when grades are changing continuously on vertical curves or when there are horizontal curves with transitions, the data sets will contain only a finite number of different values). Therefore, the following equations, which are simple applications of the definitions of the mean and standard deviation of a discrete variable taking n different values, were used to calculate the weighted average and the weighted standard deviation for the continuous

variables (i.e., on a given segment, the continuous variable is discretized by taking on only a few different values n).



a)



b)

Figure 9. A sketch illustrating the importance of considering the standard deviations a) segment with a lower standard deviation of absolute average curvature, and b) same absolute average curvature - higher standard deviation (plan view).

- Weighted average

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \quad (17)$$

- Weighted standard deviation

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n} w_i (x_i - \bar{x})^2}{\sum_{i=1}^{n} w_i}} \quad (18)$$

Where, w_i is the weight of ith sub-segment (equivalent to the length of sub-segment), and x_i is the value of an explanatory variable for ith sub-segment. Weighted averages were calculated for all continuous explanatory variables. In addition, weighted standard deviations were calculated for the degree of curvature, grade, and IRI.

Some of the roadway features (e.g., guardrails) are not installed continuously all along the roads. Therefore, to better reflect the role of explanatory variables in the model, wherever it is applicable, another set of explanatory variables were computed for the different segment lengths. For example, in the case of guardrails, the proportion of the total length of sections with guardrails in each segment was calculated to reflect its role on the frequency of RwD crashes.

In cases in which the independent variables were categorical with many classes, an attempt was made to use a dummy variable to facilitate the modeling and the interpretation of the results. Consequently, it was decided to define a variable that explains the role of these features while avoiding increasing the complexity of the model. For instance, a

dummy variable was defined as a shoulder type indicator that equals one if the shoulder type is asphalt concrete and zero otherwise.

3.3.2 The direction of the crash

As mentioned earlier in the introductory chapter, the direction of crashes may play an important role in parameter estimation of RwD crash frequency models. Hence, the direction of all RwD crashes was extracted using the information in the motor vehicle crash reports. Finding the direction of RwD to the right side was straightforward, mainly because they were mostly single-vehicle crashes. However, identifying the directions of RwD crashes to the left were a little more challenging for multi-vehicle crashes such as the head-on collisions. The direction of head-on crashes was assigned based on the direction of travel of the vehicles causing the crashes according to the police reports. Head-on collisions were seen here as simply a particular case of a RwD crash to the left side of the road with the unfortunate outcome of hitting another vehicle in the opposite direction.

RwD crashes were assigned to the segments based on the direction of travel. The geometric variables and assets data were calculated separately for each direction. Therefore, the proposed model can identify the relationship between the total number of RwD crashes in each direction and the geometric variables for that direction (not the average of both directions). Moreover, the model can consider the role of an explanatory variable like grade that can be positive or negative based on the direction of travel.

3.3.3 The general geometric environment of the roadway

The potential importance of the general geometric environment of the roadway where the section is located was discussed earlier. In order to explore its effects in the model estimation, a new set of variables (for simplicity called environmental variables) were added to the explanatory variables assuming that including additional roadway information from a certain distance upstream of each section may provide some explanatory power for modeling crashes. Therefore, the weighted average and the weighted standard deviation of curvature, grade, and IRI for 1.2-mile (~ 2.0 km) upstream of each segment were calculated and included in the models as environmental variables. As indicated before, design consistency principles indicate, for example, not only how the curvature on the analysis segment may be relevant but also how that curvature compares with the overall curvature of the preceding roadway environment. Also, in addition to the mean values, their standard deviations may be relevant as well. Therefore, the role of the general geometric environment of the roadway is indicated by adding them to the models.

3.4 Description of explanatory variables

While the definitions of some of the explanatory variables are straightforward, (e.g., lane width, shoulder width, pavement type, and shoulder type), this section explains the definitions of some other explanatory variables that are used in this research.

3.4.1 Annual Average Daily Traffic (AADT)

As mentioned earlier in section 2.2, AADT is a measure of exposure (i.e., a roadway segment with higher traffic is more prone to face RwD crashes). This effect is included in

the model by estimating the parameters as a rate model. However, in crash frequency rate analysis, it is still possible to include the AADT as an explanatory variable into the model to capture the effect that AADT may have on that rate. Therefore, in addition to including AADT with an offset (to model the rate), AADT is also included as an explanatory variable into the models. A priori, one would expect that the rate of RwD crashes would tend to decrease with AADT, as the higher traffic interactions tend to make drivers more attentive, all else equal. Since the models were developed with 10 years of data, the average over those ten years was used as the AADT for the segment.

3.4.2 The proportion of single and combination trucks in the stream

This variable is calculated by adding the proportions of single and combination trucks in the traffic stream. This variable reflects the role of heavy vehicles on the frequency of RwD crashes.

3.4.3 Mean friction demand

Pavement friction supply can be a crucial factor affecting the rate of RwD crashes as a paved road with a higher value of friction may help drivers to control their vehicles better while maneuvering on curves. However, friction supply cannot be used in this study as friction is not collected for Hawaii's state roads. Still, it was desirable to consider some measure related to friction. Thus, instead of the friction supply, which as indicated above is not currently measured, friction demand on each segment was considered.

This was motivated in part because it is relatively easy to calculate as a function of speed and because a review of the literature indicated that using the speed limit as an explanatory

variable is not usually very informative and result in counterintuitive results. For example, a study *(36)* found that the speed limit is negatively associated with the crash frequency. It was mentioned that a possible reason for this finding is that the lower speed limits are generally assigned to road segments with poor safety conditions. However, this interpretation is problematic as like in any regression model, the interpretation should be that all else equal, a road with a higher speed limit leads to lower crash rates, which is nonsensical. The problem lies in that speed is typically an endogenous variable that is affected by the same factors that affect the crash rates. Thus, it is likely correlated with the error term, which is a severe violation of the regression assumptions that may lead to biased parameter estimates (even changing its sign, as it is apparently the case with speed).

Side friction demand combines and incorporates the complex interaction between the radius of curvature, cross slope, and speed based on the horizontal curve equation from curvilinear motion. Therefore, side friction demand is included as an explanatory variable into the model to capture in part the effect of speed on RwD crashes. Equation 19 presents the relationship between these elements for a vehicle riding on a horizontal curve:

$$\frac{V^2}{15R} = f_s + e \tag{19}$$

Where V is the speed (mi/hr) (the average speed on the segment of the distress data collection van was used as an indicator of the speed of traffic on each segment), R is the radius of curvature (ft), e is the cross slope, and f_s is the side friction demand.

Therefore, side friction demand can be derived using Equation 20.

$$f_s = \frac{V^2}{15R} - e \quad (20)$$

Intuitively, it is expected that all else equal, higher friction demands would result in higher crash frequency rates.

3.4.4 Curvature

In the original database, curves to the right and curves to the left are distinguished with positive and negative signs. In this study, the curvature is calculated by taking the weighted average of the absolute value of curvature (i.e., ignoring the direction of curves). Also, to better reflect the changes in the directions of curves, using the actual values of curvatures in the database, the weighted standard deviation of the curvature is calculated for each segment.

3.4.5 Grade

A positive grade indicates that vehicles travel uphill in the segment whereas a negative grade indicates that they travel downhill. Using the directional analysis, this research employs the actual values of grades to express the differences between positive and negative grades on RwD crashes. Also, a weighted standard deviation of the grade is included in the model to identify the effect of consecutive changes in a sign of grade in a segment. It is assumed that uneven roads are associated with lower quality of riding and more chances for drives to lose control of their vehicles.

3.4.6 The International Roughness Index (IRI)

IRI (inches/mile) is an index computed from the cumulative elevation changes over a distance, as determined from a longitudinal road [1] *(37)*. It is an expression of irregularities in the pavement surface that adversely affect the ride quality of a vehicle. The a priori expectation is that a higher number of RwD crashes may occur on a segment with higher irregularities.

3.4.7 Painted median

As mentioned earlier, this research focuses on the frequency of RwD crashes on TLTW state roads. These roads are mostly undivided with no medians or physical barriers. However, the data show the existence of short segments with painted medians (usually to accommodate left-turning lanes/bays). These painted medians are supposed to provide better separation between traffic directions and accordingly, to reduce the frequency of RwD crashes. In this regard, two variables are included in the model: the width of the painted median and the proportion of the total length of sections with painted medians. Mainly, a segment with a wider painted median and a higher proportion of the total length of sections with painted medians is expected to face a fewer number of RwD crashes.

[1] It simulates the vertical movement of a quarter car with a certain combination of springs (one representing the tire and the other the suspension system) and dashpot or damper (simulating the shock absorbers).

3.4.8 Rutting

The permanent deformation of the asphalt concrete surface that accumulates in the wheel paths is referred to rutting. It is mainly the result of repeated traffic loading cycles. The expectation is that a higher value of rutting is directly associated with a higher number of RwD crashes.

3.4.9 Bridge indicator

The presence of a bridge on a segment is a common feature of TLTW state roads in Hawaii; especially for those TLTW state roads that are coastline roads (e.g., the state road 36 in Maui). This variable is included as a dummy variable in the model to identify any correlation between the existence of one or more bridges in a segment and the frequency of RwD crashes.

3.4.10 Summary of variables

Table 5 shows the description of the dummy variables. Table 6 presents the descriptive statistics of the continuous explanatory variables for the 0.2-mile segment length. Similar tables for other segment lengths are provided in Appendix B.

Table 5 Description of dummy variables

Explanatory variables	Description
Pavement type indicator	0 if it is asphalt concrete, otherwise 1
Shoulder type indicator	0 if it is asphalt concrete, otherwise 1
Bridge indicator	1 if there is a bridge in the segment, otherwise 0

Table 6 Descriptive statistics of the continuous variables

Variable	Min	Max	Mean	S.D.
AADT	545	26682	6785	5895
The Proportion of Single and Combination Trucks in the Stream	0.00	56.57	5.59	7.30
Average Side Friction Demand	0.00	0.26	0.03	0.03
Absolute Value of Curvature (degrees)	0.00	52.28	3.43	6.69
Standard Deviation of Curvature	0.00	55.24	3.59	7.25
Absolute Value of Curvature (Environmental Variable)	0.00	35.57	3.36	5.95
Standard Deviation of Curvature (Environmental Variable)	0.00	41.46	4.57	7.15
Grade (percent)	-12.89	12.07	0.00	3.11
Standard Deviation of Grade	0.00	6.68	0.49	0.66
Grade (Environmental Variable)	0.00	10.86	2.40	1.78
Standard Deviation of Grade (Environmental Variable)	0.00	7.38	1.30	0.94
IRI (inch/mile)	30.10	693.85	145.58	71.34
Standard Deviation of IRI	3.29	408.89	47.28	36.36
IRI (Environmental Variable)	34.40	586.74	144.75	65.19
Standard Deviation of IRI (Environmental Variable)	6.41	408.89	55.56	34.98
Lane Width (feet)	7.00	16.00	10.99	1.25
Painted Median Width (feet)	0.00	12.00	0.16	1.22
Rutting (inch)	0.00	0.67	0.06	0.05
Shoulder Width (feet)	0.00	18.00	4.81	3.04
The Proportion of Total Length of sections with Guardrails (mile/mile)	0.00	1.00	0.20	0.33
The Proportion of Total Length of Sections with painted Medians (mile/mile)	0.00	1.00	0.08	0.23
The Proportion of Total Length of Sections with Asphalt Concrete Shoulder (mile/mile)	0.00	1.00	0.90	0.40

CHAPTER 4: MODEL ESTIMATION

This chapter initially presents the results of the sensitivity of the estimation of model parameters on the selection of a fixed segment length for analysis. The purpose is to select a length that can be considered a compromise between segments that are too short for a meaningful analysis (too many segments with zero or one crashes) and segments that are so long that the averaging required for some explanatory variables such as curvature makes it difficult to capture the effects of some features of interest. The analysis provides some justification for the selection of a single segment length used in the subsequent analyses in this chapter. Section 4.2 presents the results (including interpretation and evaluation of each model) of three separate statistical models. Finally, section 4.3 discusses the advantages and disadvantages of each methodology and provides additional explanations regarding the model selection.

4.1 Segment length

Crash frequency histograms are helpful to visualize the effect of segment length. The crash frequency histograms for six different segment lengths (i.e., 0.1-, 0.2-, 0.3-, 0.5-, 1.0-, and 2.0-mile) are presented in Figure 10. The X-axis shows the total number of observed crashes (e.g., the possible values are 0, 1, 2, 3, …) and the Y-axis is the proportions to the total number of segments on which a given total number of crashes was observed. In addition, the figure also shows the percentages of segments with zero crashes and one crashes for each segment length. These values are calculated simply by dividing the total number of segments with zero crashes, for example, by the total number of segments.

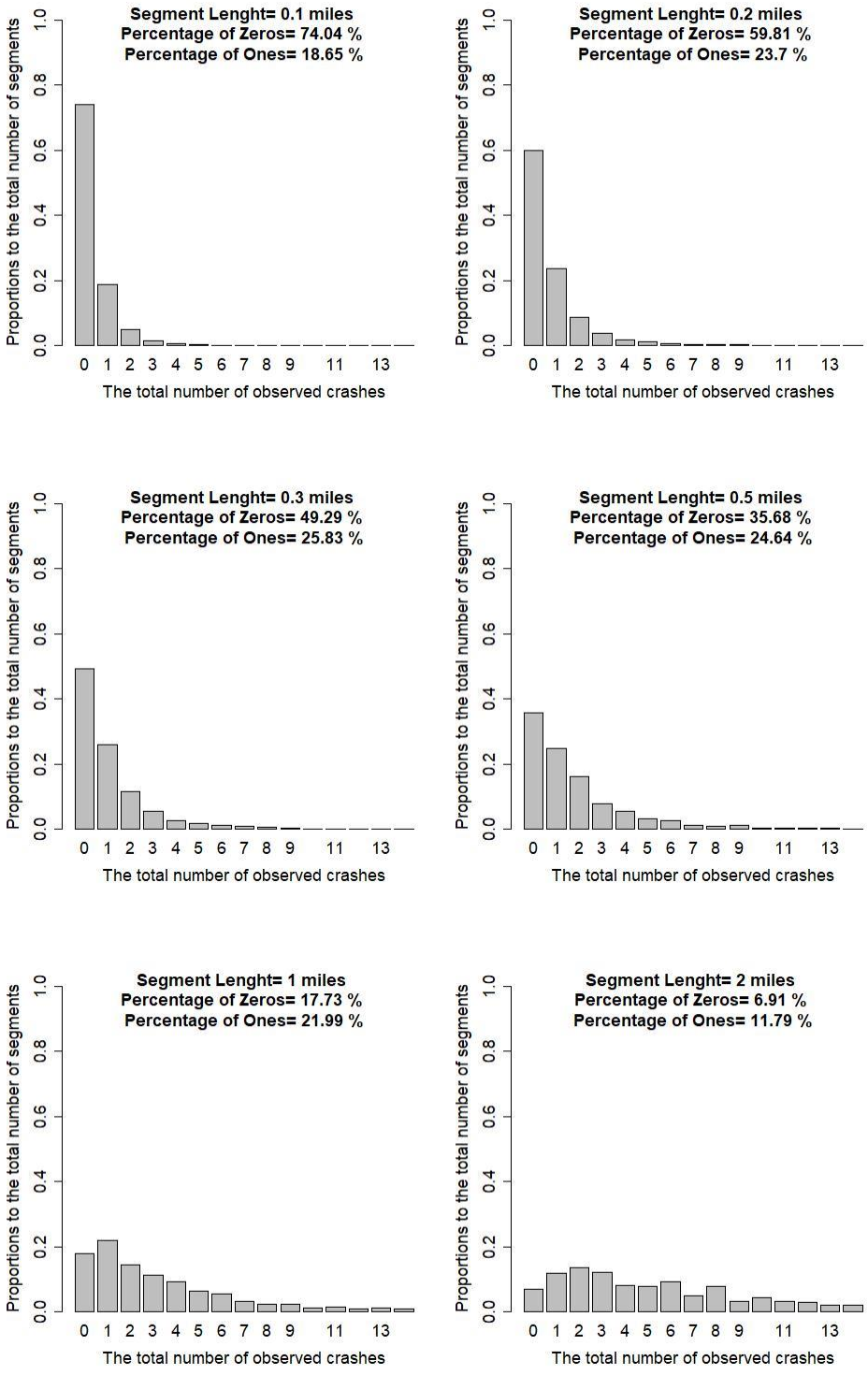


Figure 10 Crash frequency histograms for different segment lengths

As expected, the shape of crash histograms changes significantly, becoming flatter as the segment length increases. This is due to the fact that as the length increases so does the probability of occurrence of crashes even if the rate per unit length is constant; therefore, more segments are observed with a higher total number of RwD crashes. The probability of zero crashes decreases monotonically with segment length, whereas the probability of one crash in a segment increases up to segment lengths of 0.3-mile, and then decreases. The mode of the histograms also increases as the segment length increases.

As mentioned earlier in section 3.3.1, short segment lengths are more likely to result in more homogeneous segments. However, it is also essential to keep the segments long enough to maintain the purpose of crash frequency models (i.e., a model that is not developed only based on segments with zero or one crashes). In this regard, the crash frequency histogram for the 0.1-mile segment length indicates that segments with zero crashes constitute 74 percent of data. Also, it shows that about 19 percent of segments experienced only one crash (i.e., in total about 93 percent of data are segments with zero and one crash), leaving only about 7 percent of the data for identifying contributors to the frequencies of higher number of crashes. For a fixed 0.2-mile segment length, these values are approximately 59 percent and 24 percent respectively (i.e., approximately a total of 83 percent of the segments). For 0.3-mile, the same values are 49 and 26 percent, respectively; representing 75 percent of all segments.

For longer segment lengths (i.e., more than 0.3-mile) it is difficult to claim homogeneity in terms of the effects of geometric features. For example, the curvature of two segments with identical average curvature may be entirely different, since one may be

the result of a single sharp curve in an otherwise straight alignment whereas the other the combination of consecutive relatively gentle curves. This may be important for many of Hawaii's winding TLTW state routes.

The next section explores the sensitivity of the estimated models to the selection of the segment length for synthesizing the data.

4.1.1 Modeling results for different segment lengths

As discussed before, it is believed that shorter segment lengths result in more homogeneous segments. On the other hand, longer segment lengths are advantageous to define the crash frequency distribution for values above 1. In this section, the sensitivity of the model estimation to the segment length selected for analysis is explored to search for guidance for a compromise value between the homogeneity of shorter segment lengths and identification of factors contributing to higher crash frequencies.

Generally, an attempt was made to keep the segment length short enough to resist the information loss by extensive averaging. Although it is firmly believed that a segment length should not exceed more than 0.5-mile, the estimations of results for longer segment lengths are also presented to explore the effect of segment length on crash frequency models. This includes the parameter estimation of crash frequency models along with changes in the explanatory variables used for modeling the RwD crashes.

The results of six separately developed negative binomial regression models are presented in Figure 11-16. To achieve this, separate datasets including dependent and independent variables are developed for each segment length (i.e. 0.1-, 0.2-, 0.3-, 0.5-, 1.0-

, and 2.0-mile). The results include the parameter estimates, their corresponding standard errors, t-statistics, and p-values, the estimate of the dispersion parameter and its standard error, as well as the values of Akaike Information Criterion (AIC) and log-likelihood for each segment length. For each of the estimations, only the parameters with at least a 5 percent significance level are presented. The next section provides a discussion of these results.

```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)               -6.2478035  0.2515802 -24.834  < 2e-16 ***
log(AADT)                 -0.3876561  0.0271817 -14.262  < 2e-16 ***
Shoulder_width            -0.0455328  0.0075411  -6.038 1.56e-09 ***
Mean_FrictionDemand        2.4966887  0.4883375   5.113 3.18e-07 ***
Grade                     -0.0390518  0.0058903  -6.630 3.36e-11 ***
PofLengthOfMedian         -0.4407855  0.0868619  -5.075 3.88e-07 ***
PercentageTruck            0.0155059  0.0035415   4.378 1.20e-05 ***
I(Grade^2)                 0.0040136  0.0012931   3.104 0.001910 **
PofLengthOfACShoulderType -0.2296618  0.0514016  -4.468 7.90e-06 ***
Abs_Curvature              0.0885601  0.0088087  10.054  < 2e-16 ***
I(Abs_Curvature^2)        -0.0021324  0.0002757  -7.734 1.05e-14 ***
Curvature_ENV_Ave         -0.1173110  0.0226753  -5.174 2.30e-07 ***
Curvature_ENV_SD           0.0770088  0.0189129   4.072 4.67e-05 ***
PofLengthOfGuardrail      -0.2028284  0.0523000  -3.878 0.000105 ***
PaintedMedianWidth        -0.0484123  0.0193361  -2.504 0.012289 *
IRI_ENV_Ave                0.0010177  0.0003793   2.683 0.007293 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.8738) family taken to be 1)

    Null deviance: 9731.1  on 11961  degrees of freedom
Residual deviance: 8413.9  on 11946  degrees of freedom
AIC: 18605

Number of Fisher Scoring iterations: 1


              Theta:  0.8738
          Std. Err.:  0.0469

 2 x log-likelihood:  -18570.8850
```

Figure 11 Estimation results for the negative binomial regression model (segment length of 0.1-mile)

```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)               -5.8829462  0.3595570 -16.362  < 2e-16 ***
log(AADT)                 -0.3719556  0.0301789 -12.325  < 2e-16 ***
Mean_FrictionDemand        3.0029027  0.5726546   5.244 1.57e-07 ***
Shoulder_width            -0.0447175  0.0084962  -5.263 1.42e-07 ***
Grade                     -0.0416576  0.0065431  -6.367 1.93e-10 ***
PofLengthOfMedian         -0.4925268  0.1037744  -4.746 2.07e-06 ***
LaneWidth                 -0.0486305  0.0244479  -1.989  0.04669 *
PercentageTruck            0.0184675  0.0039169   4.715 2.42e-06 ***
I(Grade^2)                 0.0036652  0.0014908   2.459  0.01395 *
PofLengthOfACShoulderType -0.1890882  0.0608714  -3.106  0.00189 **
Abs_Curvature              0.0896340  0.0104201   8.602  < 2e-16 ***
I(Abs_Curvature^2)        -0.0021374  0.0003294  -6.489 8.62e-11 ***
Curvature_ENV_Ave         -0.1124697  0.0254415  -4.421 9.84e-06 ***
Curvature_ENV_SD           0.0665862  0.0212169   3.138  0.00170 **
PofLengthOfGuardrail      -0.1727271  0.0635041  -2.720  0.00653 **
PaintedMedianWidth        -0.0478326  0.0214962  -2.225  0.02607 *
IRI_ENV_Ave                0.0008950  0.0004166   2.148  0.03170 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1.1068) family taken to be 1)

    Null deviance: 6410.7  on 5963  degrees of freedom
Residual deviance: 5307.2  on 5947  degrees of freedom
AIC: 13707

Number of Fisher Scoring iterations: 1


              Theta:  1.1068
          Std. Err.:  0.0586

 2 x log-likelihood:  -13670.7540
```

Figure 12 Estimation results for the negative binomial regression model

(segment length of 0.2-mile)

```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)               -6.0640071  0.3871564 -15.663  < 2e-16 ***
log(AADT)                 -0.3527980  0.0325584 -10.836  < 2e-16 ***
Mean_FrictionDemand        2.8083184  0.6321387   4.443 8.89e-06 ***
Shoulder_width            -0.0431992  0.0092105  -4.690 2.73e-06 ***
Grade                     -0.0456889  0.0070864  -6.447 1.14e-10 ***
PofLengthOfMedian         -0.5165435  0.1180388  -4.376 1.21e-05 ***
IRI_ENV_Ave                0.0009757  0.0004438   2.198  0.02792 *
PofLengthOfGuardrail      -0.2036229  0.0749173  -2.718  0.00657 **
I(Grade^2)                 0.0039213  0.0016439   2.385  0.01706 *
I(Abs_Curvature^2)        -0.0022407  0.0003856  -5.811 6.19e-09 ***
Curvature_ENV_Ave         -0.1141678  0.0274316  -4.162 3.16e-05 ***
PercentageTruck            0.0186419  0.0042173   4.420 9.85e-06 ***
PofLengthOfACShoulderType -0.1489750  0.0642626  -2.318  0.02044 *
Abs_Curvature              0.0916615  0.0118639   7.726 1.11e-14 ***
Curvature_ENV_SD           0.0687404  0.0228440   3.009  0.00262 **
PaintedMedianWidth        -0.0540200  0.0235862  -2.290  0.02200 *
LaneWidth                 -0.0516936  0.0261239  -1.979  0.04784 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1.2421) family taken to be 1)

    Null deviance: 4840.6  on 3966  degrees of freedom
Residual deviance: 3878.5  on 3950  degrees of freedom
AIC: 11196

Number of Fisher Scoring iterations: 1


              Theta:  1.2421
          Std. Err.:  0.0669

 2 x log-likelihood:  -11160.2620
```

Figure 13 Estimation results for the negative binomial regression model (segment length of 0.3-mile)

```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)               -6.7918249  0.3316009 -20.482  < 2e-16 ***
log(AADT)                 -0.3316668  0.0362439  -9.151  < 2e-16 ***
Shoulder_width            -0.0467074  0.0103454  -4.515 6.34e-06 ***
Mean_FrictionDemand        2.5260766  0.7302731   3.459 0.000542 ***
Grade                     -0.0497586  0.0080107  -6.212 5.25e-10 ***
PofLengthOfMedian         -0.6996651  0.1479947  -4.728 2.27e-06 ***
IRI_ENV_Ave                0.0014097  0.0005025   2.805 0.005025 **
PofLengthOfGuardrail      -0.2013805  0.0876129  -2.299 0.021532 *
I(Grade^2)                 0.0039452  0.0019339   2.040 0.041352 *
PofLengthOfACShoulderType -0.2195588  0.0668502  -3.284 0.001022 **
PercentageTruck            0.0172567  0.0045856   3.763 0.000168 ***
PaintedMedianWidth        -0.0683425  0.0285515  -2.394 0.016681 *
Curvature_ENV_Ave         -0.1020939  0.0317426  -3.216 0.001299 **
Abs_Curvature              0.1062226  0.0137506   7.725 1.12e-14 ***
I(Abs_Curvature^2)        -0.0027644  0.0004610  -5.997 2.01e-09 ***
Curvature_ENV_SD           0.0544675  0.0260844   2.088 0.036787 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1.4652) family taken to be 1)

    Null deviance: 3305.6  on 2372  degrees of freedom
Residual deviance: 2509.3  on 2357  degrees of freedom
AIC: 8425.6

Number of Fisher Scoring iterations: 1


              Theta:  1.4652
          Std. Err.:  0.0844

 2 x log-likelihood:  -8391.6060
```

Figure 14 Estimation results for the negative binomial regression model (segment length of 0.5-mile)

```
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -3.548060   0.485405  -7.309 2.68e-13 ***
log(AADT)             0.635514   0.041218  15.418  < 2e-16 ***
Mean_FrictionDemand   6.469875   0.947698   6.827 8.67e-12 ***
PofLengthOfMedian    -1.109586   0.191336  -5.799 6.67e-09 ***
Grade                -0.056329   0.009861  -5.712 1.11e-08 ***
Shoulder_width       -0.042457   0.012665  -3.352 0.000801 ***
IRI_ENV_Ave           0.001553   0.000588   2.642 0.008252 **
PofLengthOfGuardrail -0.244548   0.110749  -2.208 0.027235 *
PercentageTruck       0.015676   0.005420   2.892 0.003824 **
LaneWidth            -0.078398   0.032902  -2.383 0.017183 *
Curvature_ENV_Ave    -0.048637   0.010897  -4.463 8.07e-06 ***
Curvature_SD          0.142826   0.027932   5.113 3.16e-07 ***
Abs_Curvature        -0.130376   0.033300  -3.915 9.03e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1.8574) family taken to be 1)

    Null deviance: 1736.9  on 1168  degrees of freedom
Residual deviance: 1268.1  on 1156  degrees of freedom
AIC: 5419.9

Number of Fisher Scoring iterations: 1


              Theta:  1.857
          Std. Err.:  0.122

 2 x log-likelihood:  -5391.864
```

Figure 15 Estimation results for the negative binomial regression model (segment length 1.0-mile)

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6389  -0.9409  -0.2391   0.4142   4.5130

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               -6.776946   0.406101 -16.688  < 2e-16 ***
log(AADT)                 -0.308730   0.049469  -6.241 4.35e-10 ***
Mean_FrictionDemand        8.188813   1.209948   6.768 1.31e-11 ***
PofLengthOfMedian         -1.426041   0.291508  -4.892 9.98e-07 ***
Grade                     -0.052396   0.012305  -4.258 2.06e-05 ***
PofLengthOfACShoulderType -0.295700   0.091935  -3.216 0.001298 **
I(Grade^2)                 0.009003   0.003165   2.844 0.004451 **
PaintedMedianWidth        -0.145920   0.048845  -2.987 0.002813 **
PercentageTruck            0.022642   0.005957   3.801 0.000144 ***
Shoulder_width            -0.058706   0.014599  -4.021 5.79e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(2.3782) family taken to be 1)

    Null deviance: 1040.46  on 572  degrees of freedom
Residual deviance:  616.53  on 563  degrees of freedom
AIC: 3318.8

Number of Fisher Scoring iterations: 1


              Theta:  2.378
          Std. Err.:  0.192

 2 x log-likelihood:  -3296.812
```

Figure 16 Estimation results for the negative binomial regression model (segment length 2.0-mile)

4.1.2 Discussion

In all the models, the variables with statistically significant parameter estimates (i.e., those identified as contributing factors of RwD crashes) and their signs are intuitively correct. The detailed discussion of the significance of individual model parameters and their interpretations is postponed until Section 4.2. This section discusses the main differences between the models that are developed with different segment lengths.

The results show that the total number of statistically significant parameters changes with the selected segment length. The numbers of significant parameters are 16, 17, 17, 16, 13, and 10 for 0.1-, 0.2-, 0.3-, 0.5-, 1.0-, and 2.0-mile segment lengths, respectively. The main difference between the models for 0.1-mile and 0.2-mile segment lengths is that the effect of lane width is statistically significant for the 0.2-mile model (though very close to the 5 percent significance level) but not for the 0.1-mile model. This may be a consequence of the relatively small percentage of segments with observations of more than one crash for the 0.1-mile segment length. Otherwise, the parameter estimates are of similar magnitudes, albeit, with generally smaller absolute values for the 0.2-mile segment length.

The comparison between the 0.2-mile and 0.3-mile segment lengths is very similar, with the magnitude of some parameter estimates increasing and of others decreasing with the segment length. When the segment length used for analysis is increased to 0.5-mile, the lane width becomes again non-significant at the 5 percent level. Increasing the segment length further to 1.0-mile and 2.0-mile result in the loss of 3 statistically significant parameters in each case. This is to be expected, as the range of some of the explanatory

variables decreases by increasing the segment length, resulting in information loss caused by averaging the information. When changing from 0.5-mile to 1.0-mile segment length, the variables $Grade^2$ and $Abs\_Curvature^2$ become non-significant. These variables represent second-order effects that become more difficult to identify with longer segment lengths. Also, the variable Curvature_ENV_SD (the standard deviation of curvature in a 1.2-mile segment prior to the analysis segment) becomes non-significant whereas the variable Curvature_SD becomes significant. This finding, which indicates that the standard deviation of curvature prior to the analysis segment is less informative, is reasonable because by increasing the segment length the role of environmental variables diminishes.

When changing from 1.0-mile to 2.0-mile segment length, the variables IRI_ENV_AVG, PofLengthOfGuardrail, ABS_Curvature, Curvature_SD and Curvature_ENV_Ave which appear in the 1.0-mile model are not present in the 2.0-mile model. On the other hand, the variables PofLengthOfACShoulderType and PaintedMedianWidth on the 2.0-model are missing on the 1.0-model. The fact that the parameters of curvature related variables become insignificant for the longer segment lengths is reasonable because the range of the curvature related variables will necessarily decrease with segment length, and they will not be informative to the model. The disappearance of the proportion of the total length of sections with guardrails can also be explained by the fact that there are not many sections with a long segment of guardrails[1]. Thus, as the segment length becomes larger, the

[1] Generally, guardrails are expensive countermeasures, and economic feasibility may prevent them from being placed everywhere.

differences captured by this variable become smaller. Also, with very long segment lengths, the average roughness of the segment prior to the analysis segment is of little importance (unless a crash occurs at the beginning of such long segments, it makes sense that the roughness of the prior segment would not be important).

The environmental variables (those capturing the effect of the alignment prior to the analysis segment) are not statistically significant in the model with the 2.0-mile segment length. This is intuitively correct since for long segments the features of the prior segment probably have an effect only over a short initial portion of the analysis segment.

It is also observed that there is a monotonic increase of the estimate of the dispersion parameter and its standard error with segment length. Consequently, it appears that in all situations, the assumption that the data generation process for crash frequencies can be represented by a single Poisson distribution is inadequate but that this assumption gets even more untenable for longer segment lengths. This relates to another concern with the usual practice of defining "homogeneous" segments of widely different lengths for model estimation. As illustrated from the results in this section, this practice may result in segments for which the data generation processes follow distributions with very different dispersion parameters. Besides, although as pointed out in chapter 3 the homoscedasticity assumption is usually relaxed in GLMs, the use of widely different segment lengths for estimation of a single model would tend to make the variability of the longer segments artificially smaller than the variability of the smaller segments. In other words, in addition to the natural heteroskedasticity of the data synthesis process, the use of different segment

lengths in a given analysis would tend to add artificial heteroscedasticity that if not properly accounted for would result in less efficient parameter estimates.

In conclusion, the results show that the segment length affects which variables enter the crash frequency models. Although it cannot be concluded from this type of analysis what segment length is the most adequate, it is seen that similar models are obtained with segment lengths between 0.2- mile to 0.3-mile or even with 0.5-mile. Under the premise that shorter segment lengths result in more homogeneous segments and with the inclusion of environmental variables that incorporate the effects of the road environment prior to the segment, the following analyses in this dissertation are provided only for the 0.2-mile segment length. As shown before, the 0.1-mile segment length results in a tiny proportion of segments with more than one crash over the study period, which was not considered desirable. Also, a larger segment length (up to about 0.5-mile) produced qualitatively similar results. Thus, similar conclusions would be reached with other lengths. It is believed that with larger segment lengths the identification of important factors becomes more difficult.

4.2 Crash frequency models

In this section, various GLM are employed to develop crash frequency models. Amongst the existing methodologies, the negative binomial regression model, zero-inflated negative binomial regression model, and the mixed-effects negative binomial regression model are used. The results of each approach are presented separately by considering the segment length equal to 0.2-mile. The evaluations of these methodologies

including a discussion on the advantages and disadvantages of each methodology are provided in section 4.3.

4.2.1 The negative binomial regression model

The results of the negative binomial regression model, estimated using the R-package Mass *(38)*, are presented in Figure 17. The parameters of the negative binomial regression model are estimated using maximum likelihood estimation. Figure 17 presents the parameter estimates, their corresponding standard errors, p-values, the estimated dispersion parameter and its standard error, the Akaike Information Criterion (AIC), and the value of $2 \times$ maximum log-likelihood function. A detailed interpretation of the results is provided in the next section.

The correlations between explanatory variables were investigated prior to the modeling to reduce the issues of efficiency caused by multicollinearity or to avoid the simultaneous inclusion of highly correlated variables. For example, the absolute value of curvature is highly correlated ($\rho = 0.98$) with the standard deviation of curvature, hence only the absolute value of curvature is included in the model.

4.2.1.1 Interpretation of results

In this section, the results of the negative binomial regression model are explained. This includes the interpretation of statistically significant parameters and the model's statistics.

```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -5.8829462  0.3595570 -16.362  < 2e-16 ***
log(AADT)                  -0.3719556  0.0301789 -12.325  < 2e-16 ***
Mean_FrictionDemand         3.0029027  0.5726546   5.244 1.57e-07 ***
Shoulder_width             -0.0447175  0.0084962  -5.263 1.42e-07 ***
Grade                      -0.0416576  0.0065431  -6.367 1.93e-10 ***
PofLengthOfMedian          -0.4925268  0.1037744  -4.746 2.07e-06 ***
LaneWidth                  -0.0486305  0.0244479  -1.989  0.04669 *
PercentageTruck             0.0184675  0.0039169   4.715 2.42e-06 ***
I(Grade^2)                  0.0036652  0.0014908   2.459  0.01395 *
PofLengthOfACShoulderType  -0.1890882  0.0608714  -3.106  0.00189 **
Abs_Curvature               0.0896340  0.0104201   8.602  < 2e-16 ***
I(Abs_Curvature^2)         -0.0021374  0.0003294  -6.489 8.62e-11 ***
Curvature_ENV_Ave          -0.1124697  0.0254415  -4.421 9.84e-06 ***
Curvature_ENV_SD            0.0665862  0.0212169   3.138  0.00170 **
PofLengthOfGuardrail       -0.1727271  0.0635041  -2.720  0.00653 **
PaintedMedianWidth         -0.0478326  0.0214962  -2.225  0.02607 *
IRI_ENV_Ave                 0.0008950  0.0004166   2.148  0.03170 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1.1068) family taken to be 1)

    Null deviance: 6410.7  on 5963  degrees of freedom
Residual deviance: 5307.2  on 5947  degrees of freedom
AIC: 13707

Number of Fisher Scoring iterations: 1


              Theta:  1.1068
          Std. Err.:  0.0586

 2 x log-likelihood:  -13670.7540
```

Figure 17 Estimation results for the negative binomial regression model

4.2.1.1.1 logarithm (AADT)

Generally, AADT should be considered as a measure of exposure to the frequency RwD crashes since a roadway segment with higher traffic flow is prone to face more RwD crashes. As explained in section 2.2 titled “Rate Models,” this is typically accomplished by including AADT as an offset. However, in crash frequency rate analysis, even with the use of an offset for AADT it is still possible to include the AADT as an explanatory variable

in the model. Note that with the use of an offset, the parameter estimate for this variable is equivalent to (β_1-*1)*. The math behind this coefficient is explained below.

As explained earlier in section 2.1.2, the parameters of the negative binomial regression model are estimated by a log link function such as Equation 21 as:

$$\lambda_i = exp(\boldsymbol{\beta x}_i + \varepsilon_i) \tag{21}$$

Which is equivalent to:

$$\lambda_i = e^{\beta_0 + \beta_1 (Variable\ 1) + \beta_2 (Variable\ 2) + \ldots + \beta_n (Variable\ N) + \varepsilon_i} \tag{22}$$

If *log(AADT)* is inserted instead of *Variable1*, then:

$$\lambda_i = e^{\beta_0 + \beta_1 \log(AADT) + \beta_2 (Variable\ 2) + \ldots + \beta_n (Variable\ N) + \varepsilon_i} \tag{23}$$

Meanwhile, as discussed in section 2.2, the intercept in a rate model with an offset for log(AADT) is $\beta_0 = Intecept + offset(log(AADT))$, where offset means that the variable is entered into the model with a parameter equal to 1 (other offsets are treated similarly); hence, when log(AADT) is specified as an explanatory variable but also included as an offset, Equation 23 above turns into:

$$\lambda_i = e^{\beta_0 + (\beta_1 - 1)(log(AADT)) + \ldots + \beta_n (Variable\ N)} \tag{24}$$

where β_1 is the model parameter that would have been estimated if log(AADT) had been specified as an explanatory variable but without any offset for log(AADT). With an offset, the parameter estimate for log(AADT) becomes $\beta_1^{'} = \beta_1 - 1$ to compensate for the parameter of log(AADT) being fixed at 1 within β_0. The advantage of including an offset

is that it allows the assessment of the effect of log(AADT) on the average rate of crashes (i.e., on $\lambda_i/AADT$).

The estimate of the parameter for the logarithm of AADT is statistically significant (p-value < 2×10^{-16}). The negative sign of this parameter implies that a segment with a higher traffic volume is expected to have a lower rate of RwD crashes. This finding may be explained by the fact that a higher traffic volume induces drivers to drive more attentively because of the higher chances of incidents with other vehicles.

4.2.1.1.2 The proportion of single and combination trucks in the traffic stream

The proportion of single and combination trucks in the traffic stream is found to be statistically significant in the model (p-value = 2.42×10^{-6}). The positive coefficient implies an increase in the frequency of RwD crashes with increasing trucks percentages. The tendency of the drivers to overtake the trucks while driving behind them, usually because of the lowers average speed of trucks and the desire to have more visibility of the roadway ahead, may increase the chance of being involved in RwD crashes[1].

4.2.1.1.3 Mean side friction demand

The mean side friction demand variable and its calculation are explained in section 3.4.3. The parameter estimation of the negative binomial model confirms the importance of this variable to the frequency of RwD crashes by finding it statistically significant (p-value = 1.57×10^{-7}). The positive sign of this coefficient suggests that a roadway

[1] More specifically being involved in RwD crashes to the left side, also known as head-on crashes in literature.

segment with a higher mean side friction demand[1] is more likely to experience a higher number of RwD crashes. This is considered a very reasonable result as it is easier for drivers to lose control of their vehicles on segments requiring higher mean side friction demand. Note that this variable should not be confused with the maximum friction that could be supplied at the interface of the road surface and the vehicle tires. The latter is influenced by other factors such as the weather condition, the tread depth of tires, and the driving speed. Because of its importance for safety, it would have been desirable also to include friction supply; however, HDOT does not currently perform pavement friction measurements.

4.2.1.1.4 The average of the absolute value of the curvature

In this research, a quadratic function[2] is assumed for the absolute value of curvature. Thus, two parameters are estimated for this variable. Both parameters are statistically different from zero with a p-value of less than 2×10^{-16} for the parameter of the linear term and a p-value of 8.62×10^{-7} for the parameter of the quadratic term. Figure 18 illustrates the resulting contribution of the average of the absolute value of curvature to the link function in the negative binomial model. As expected, the graph is a parabola, depicting the variability in the magnitude of this contribution. Except for very extreme high average curvatures, the estimated contribution is generally positive for the available range

[1] It derives from the interaction of radius of curvature, side slope and the speed.

[2] The general form of a quadratic function is $\beta_1\ (Variable)+\beta_2(Variable)^2$. Note that a constant term is not needed since this is absorbed within the constant term of the linear function.

of curvatures in the dataset. However, the magnitude of this contribution changes in this range and its slope changes sign.

For the segments with curvatures less than twenty degrees, the estimated contribution is increasing at a decreasing rate. This means that all else equal, a higher rate of crashes should be expected on segments with any positive curvature below 20 degrees than on a straight segment. Furthermore, the closer the curvature is to 20 degrees, the higher is the expected rate of crashes. It must be noted that most segments fall in this range (i.e., 96 percent). On the other hand, for segments with curvatures more than twenty degrees, the estimated contribution is decreasing in the magnitude at an increasing rate. A possible explanation is that drivers tend to start paying more attention while driving on segments with a high degree of curvature, which may partially neutralize the adverse effect of curvature on the frequency of RwD crashes. Note that for some very extreme curvatures, the estimated effect becomes negative. These represent very extreme situations where driving may be so dangerous that drivers take extra precaution while driving.

The positive contribution of curvature implies a higher likelihood of RwD crashes on curvy roads. This is consistent with the findings of a recent study that implies the degree of curvature is positively associated with the frequency of head-on crashes *(39)*. Moreover, this result is intuitively reasonable since the sight distance and visibility are notably limited in curvy roads.

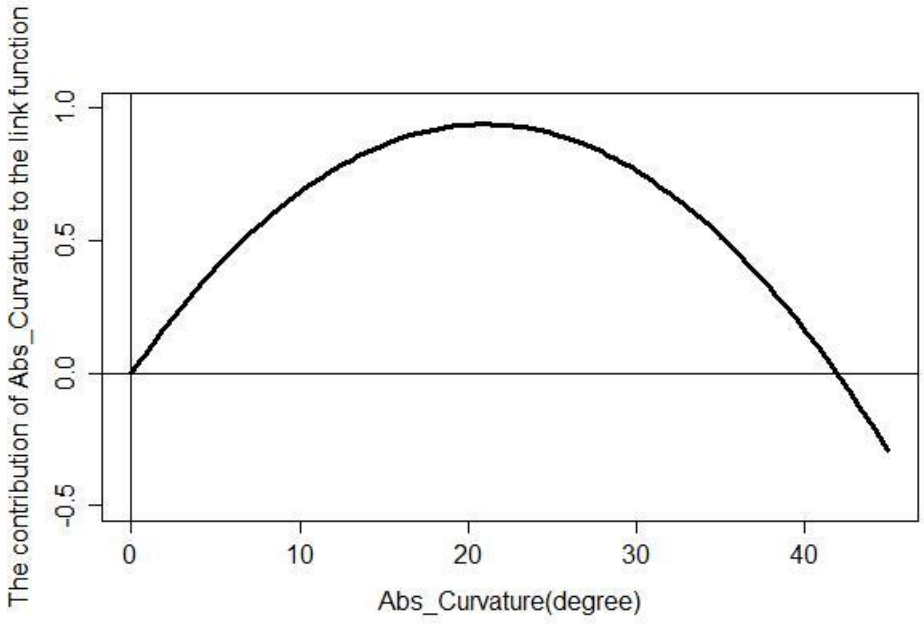


Figure 18 The contribution of curvature to the link function (quadratic function)

4.2.1.1.5 Grade

A quadratic function is also assumed for this variable to explore its effects on the frequency of RwD crashes. Both parameters are statistically significant with a p-value of 1.93×10^{-10} for the linear term and a p-value of 0.014 for the quadratic term. The directional analysis of RwD crashes enabled the model to capture better the effect of grade that can take both positive and negative values (whether the segment is in an uphill or in a downhill, respectively).

Figure 19 presents the contribution of grade to the link function in the negative binomial model. This graph is drawn for the available range of grades in the dataset. For the negative grades, it shows that the estimated contribution to the link function is positive at an increasing rate as the grade becomes more negative. This means that the frequency of

RwD crashes tends to increase for steeper downgrades and that the rate of increase in the frequency of RwD crashes gets higher for segments on steeper downgrades. This result is intuitively correct since it is easier for drivers, particularly of heavy vehicles, to lose control of their vehicles while going downhill.

On the other hand, the estimated contribution to the link function is generally negative for positive grades (uphill). This is seen on the right side of Figure 19. This indicates that all else equal, segments with average positive grades tend to be slightly safer than segments with average zero grade. All else equal, upgrades of about 6 percent (where the minimum contribution is observed) tend to have the lowest rate of RwD crashes. It must be noted that about 93 percent of segments fall between -6 percent and +6 percent grade. A possible reason for this result is that positive grades help to reduce the speed in emergency situations, thus helping to reduce the frequency of RwD crashes. In general, the magnitudes of the contribution to the link function are substantially smaller for the positive grades in comparison to the negative grades. This finding indicates that negative grades are more crucial for studying the frequency of RwD crashes.

4.2.1.1.6 Lane width

The parameter for lane width (-0.0486) is statistically significant at a 5 percent significance level (p-value = 0.0467) with a negative sign. This result is intuitively valid because wider traffic lanes are associated with better traffic separation in each direction. Also, wider traffic lanes provide more space for drivers to maneuver, and it may alleviate traffic conflicts.

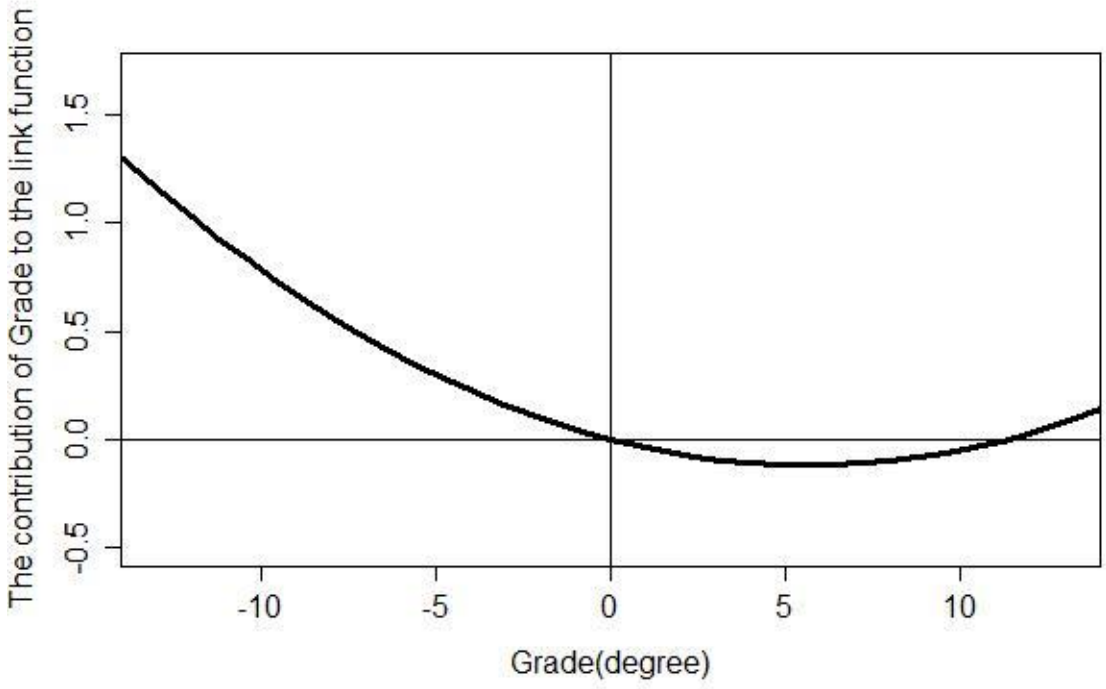


Figure 19 The contribution of grade to the link function (quadratic function)

4.2.1.1.7 Shoulder width

The estimate of the parameter for shoulder width is also negative (-0.0447), implying a reduction in the frequency of RwD crashes with increasing width. The p-value is 1.42×10^{-7}. Wider shoulders provide additional recovery space for drivers to control the vehicles, especially in the run-off the road crashes. This result is in agreement with the findings presented in the literature *(12).*

4.2.1.1.8 Painted median width

The negative parameter estimate for painted median width (-0.0478) indicates that the segments with a higher painted median width face fewer RwD crashes. The

corresponding p-value is 0.026. A possible explanation is that a wider separation between each traffic direction effectively diminishes the chance of RwD crashes to the left side including the head-on crashes.

4.2.1.1.9 The proportion of the total length of the sections with painted medians

The estimated parameter for this variable (-0.4925) is negative with a p-value = 2.07×10^{-6}. This finding reveals that a segment with a higher proportion of the total length sections with painted medians faces fewer RwD crashes. A higher proportion of the total length of sections with painted medians provides better traffic separations in each direction and can effectively reduce the chances of head-on collisions.

4.2.1.1.10 The proportion of the total length of sections with guardrails

The estimate of this parameter is also negative (-0.1727), indicating that a higher proportion of the total length of the guardrails is associated with a fewer number of RwD crashes. The p-value is 0.0063. The existence of guardrails, which can be viewed as a continuous obstacle reducing the critical rate of the visual angle, may cause drivers to pay more attention to the changes in roadway geometry that could contribute to a lower chance of being involved in a RwD crash.

4.2.1.1.11 The absolute value of curvature (environmental variable)

In section 3.3.3, it was explained how the absolute value of curvature (environmental variable) could affect the frequency of RwD crashes. In this dissertation, the curvature of a segment and the curvature for 1.2-mile before each segment

(environmental variable) are included separately in the model to evaluate the importance of design consistency in highway design practices.

In addition to the effects of the absolute value of curvature in the segment described before, the estimation results identified the absolute value of curvature (environmental variable) as statistically significant (p-value = 9.84×10^{-6}). Interestingly, unlike the absolute value of curvature, the estimated parameter of the environmental variable (-0.1125) has a negative sign, indicating a reduction of the frequency of RwD crashes with increasing values of the average environmental curvature before each segment. Although this finding may be surprising at first, it has a very intuitive rational explanation. With all else equal, including the average curvature on a segment, the rate of RwD departure crashes decreases with higher curvatures of the alignment prior to the segment. This finding means that if a driver is riding a car on a uniformly winding road, there would be less chance to get involved in RwD crashes. This might be referred to a higher preparedness of the drivers riding on a uniformly winding road either by reducing their speeds and/or by paying more attention to the geometry of the road ahead. Therefore, this finding highlights the importance of design consistency in highway design and its safety benefits.

4.2.1.1.12 The standard deviation of curvature (environmental variable)

The parameter estimate of the standard deviation of curvature (environmental variable) is positive (0.0666) and statistically significant (p-value = 0.0017). This result also agrees with design consistency concepts. Although a higher average curvature prior to a given segment may lead to lower crash rates, the distribution of that curvature is also important. For a given average curvature, roads with gradual changes in curvature tend to

be safer than roads composed of long tangents and sharp curves. This means that there is a higher chance of getting involved in a RwD crash if a substantial geometric variation (i.e., the curvature) exists before each segment on the road. The parameters for two variables related to environmental curvature together with the variables relating to the curvature of the segment seem to work together to capture design consistency issues.

4.2.1.1.13 IRI (environmental variable)

The parameter estimate for the IRI on the segment prior to the analysis segment (0.000895) was found to be statistically significant (p-value = 0.0317). This variable measures the roughness of the pavement surface. A higher IRI value for a segment indicates a lower ride quality on that segment. The IRI value is computed by simulating the up and down movement of a quarter-car of standardized characteristics per unit distance traveled, and it is measured in inches/mile or mm/km. In this study, the data was used in inches/mile. The above estimation result implies that the pavement irregularities before each segment increases the frequency of RwD crashes. Interestingly, the roughness on the segment was not found to be statistically significant.

4.2.1.2 Model evaluation

This section provides an analysis to evaluate the goodness of fit of the estimated negative binomial regression model. The literature on crash frequency models usually presents the “best model” selected based on some criterion such as the log-likelihood, the AIC, BIC. However, while these are useful to select a model, they are difficult to interpret.

Unfortunately, unlike a regression model for a continuous dependent variable, a plot of deviances (the equivalent of deviance residuals for this type of models) is not very useful either. Figure 20 shows the deviance residuals for the model presented in this section. The bulk of the data corresponds to zero, one, or two crashes (darker areas), and the density of points decreases for more crashes since there are fewer observations.

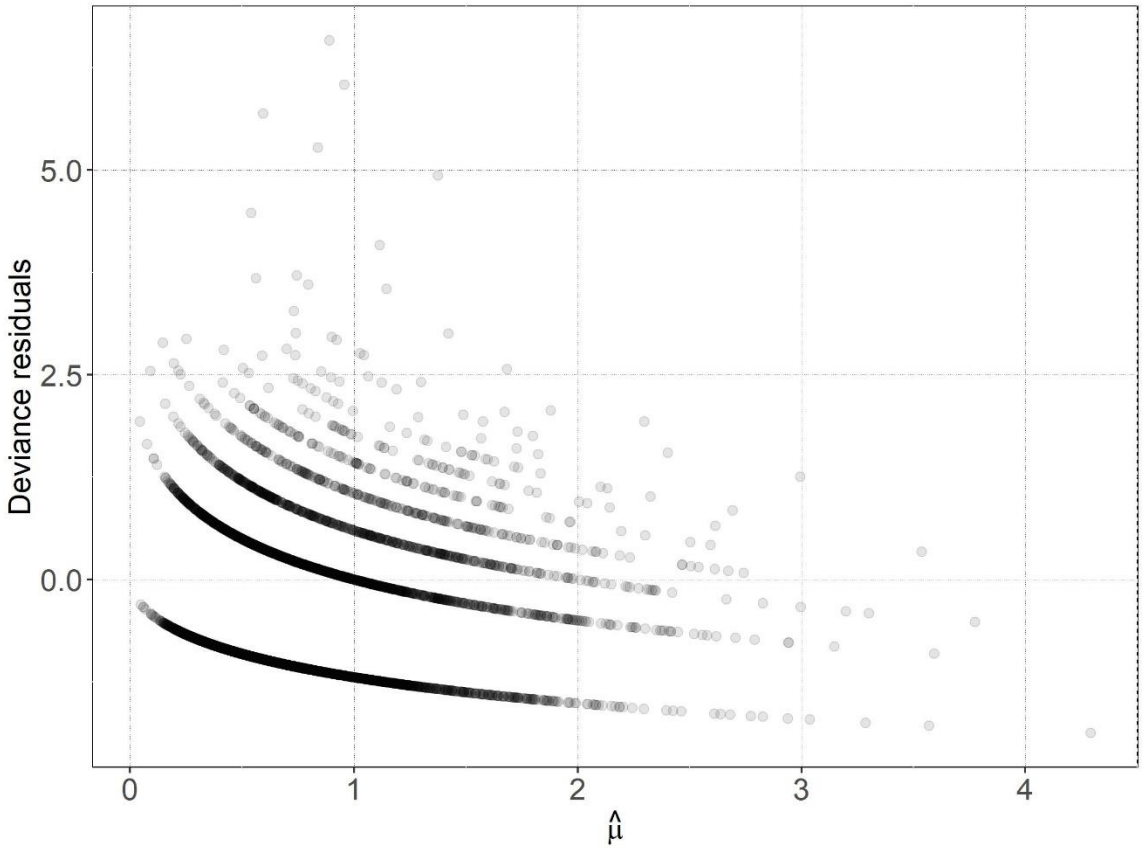


Figure 20 Deviance residuals versus $\hat{\mu}$

Some researchers *(16)* present a plot of estimated vs. observed values. This is considered misleading since what is presented as observed values is the actual number of crashes observed on a given section, which is a particular realization of the random number

of crashes that could have occurred in that section. On the other hand, what is presented as predicted is typically the predicted average number of crashes. Clearly, these two variables are not the same. This is also evidenced in what appears as systematic bias in the predictions, as shown in Figure 21 for the model presented in this section. The problem with this approach is that in addition to comparing two different variables, it fails to recognize the probabilistic nature of the model being estimated. For this reason, this section presents a novel approach to evaluate the goodness of fit of the model graphically. The idea behind the plots is consistent with the probabilistic nature of crash frequency models.

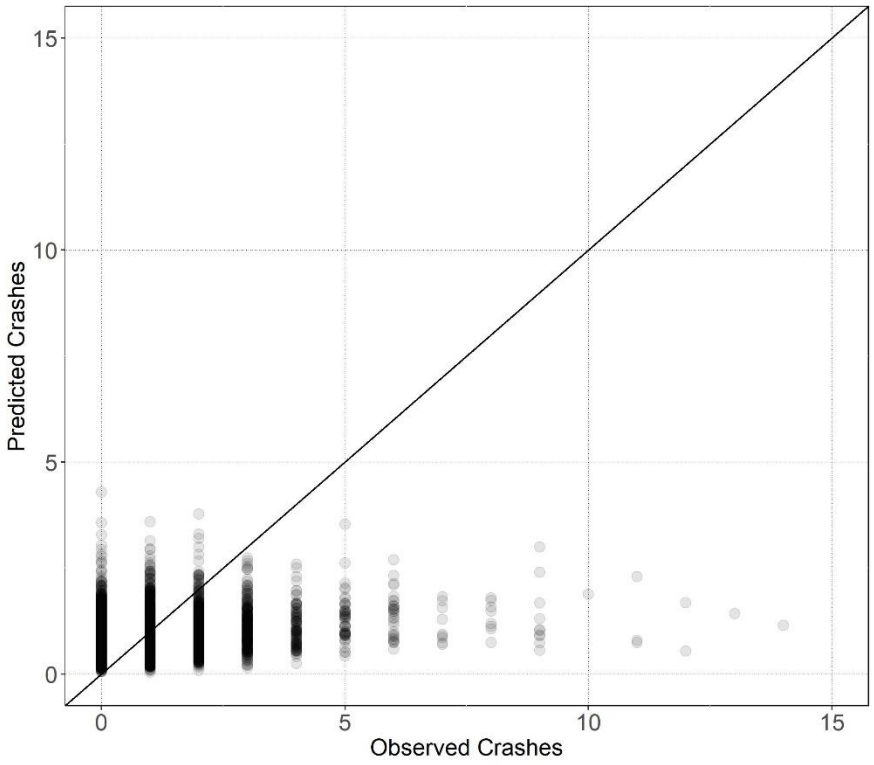


Figure 21 Observed versus predicted crashes

Although crash frequency models are necessarily developed from observational studies where the researcher has no control over the independent variables, it is useful to

think about a hypothetical experiment with a large number of identical sections (such as the same curvature, traffic, grades). In such an experiment, no crashes would be observed on a certain proportion of the sections; another set of sections would have only one crash, another would have two crashes and so on. If the estimated model provides a good representation of the data generation process, then one would expect to have similar observed vs. predicted proportions (or probabilities). The problem is that the data in this study represent a cross-section of data, with a single observation for each section. Thus, in order to have replicates one would need to identify nearly identical sections with some kind of clustering algorithm. With so many explanatory variables, this becomes difficult even with a large dataset.

Instead, in this study, the estimated model was used to predict the average number of crashes for each section and then the data were segmented based on this prediction. Sections were assigned to the different data segments defined by narrow intervals for the predicted average crash frequency (e.g., 0.0-0.2, 0.2-0.4, 0.4-0.6, etc.). Now, for a given data segment, the mid-value of the range was used to predict the probabilities of 0, 1, 2, … crashes in the segment. For example, for the data segment defined by a predicted average of crashes between 0.4-0.6, the value of $\lambda = 0.5$ was used to predict the probabilities. Those probabilities were then compared with the proportion of the sections assigned to the data segment on which 0, 1, 2, … crashes occurred.

Figures 22-31 compare the observed distributions (i.e., the observed proportions of sections with 0, 1, 2, … RwD crashes) with the distributions generated with negative binomial probability predictions for each data segment. The similarity between the

distribution of the data and the negative binomial distribution can confirm the validity of the estimation. It is worth mentioning again here that the parameters of the negative binomial regression model are estimated by using the maximum likelihood approach, and by finding the probability of observing a certain number of crashes for each segment based on the negative binomial distribution.

The shape of the negative binomial distribution can vary considerably for different ranges of its parameters (i.e., the average and the dispersion parameter). This is why it is essential to define relatively narrow ranges for the data segmentation.

The total number of observations (i.e., the total number of segments whose expected number of RwD crashes falls within the corresponding range for each figure) is provided on top of each graph. This value varies considerably for different ranges. Note that the estimation is supposed to be more accurate if the distribution is drawn based on a considerable number of observations.

The red bars on each figure represent the probability distribution of the number of RwD crashes generated with the segments for which the predicted average number of crashes using the estimated model parameters of the negative binomial distribution fall within the narrow range for each figure. It is worth repeating that these are simply the proportions of the segments selected for each figure that have 0, 1, 2, … number of crashes. On the other hand, the blue bars illustrate the negative binomial probability distribution predictions using the mid-value of the range of the average number of crashes corresponding to each figure. These are drawn with a dispersion parameter equal to the estimated dispersion parameter of the negative binomial model.

In general, the observed and predicted distributions are remarkably similar. Figures 22-31 exhibit acceptable similarities between the two probability distributions especially for the ranges with more than 100 observations. There is a slight underprediction of the proportions of zero outcomes for the two ranges with the most data (0.4-0.6 and 0.6-0.8), which represent 45 percent of the plotted data which is compensated with overpredictions in the other ranges.

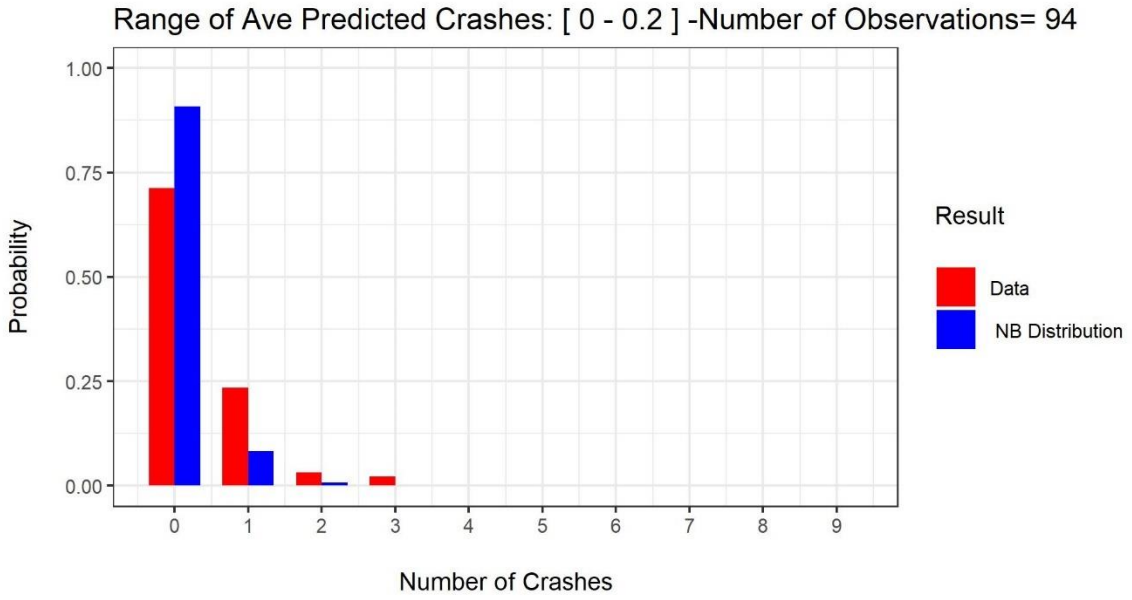


Figure 22 Comparison graph for the first range of average predicted crashes

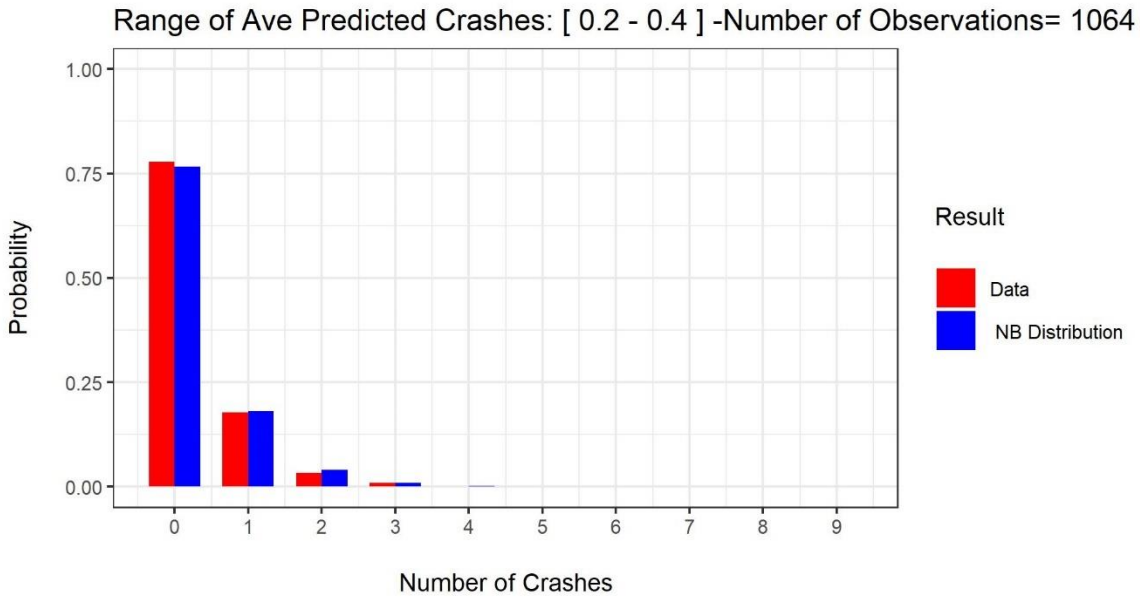


Figure 23 Comparison graph for the second range of average predicted crashes

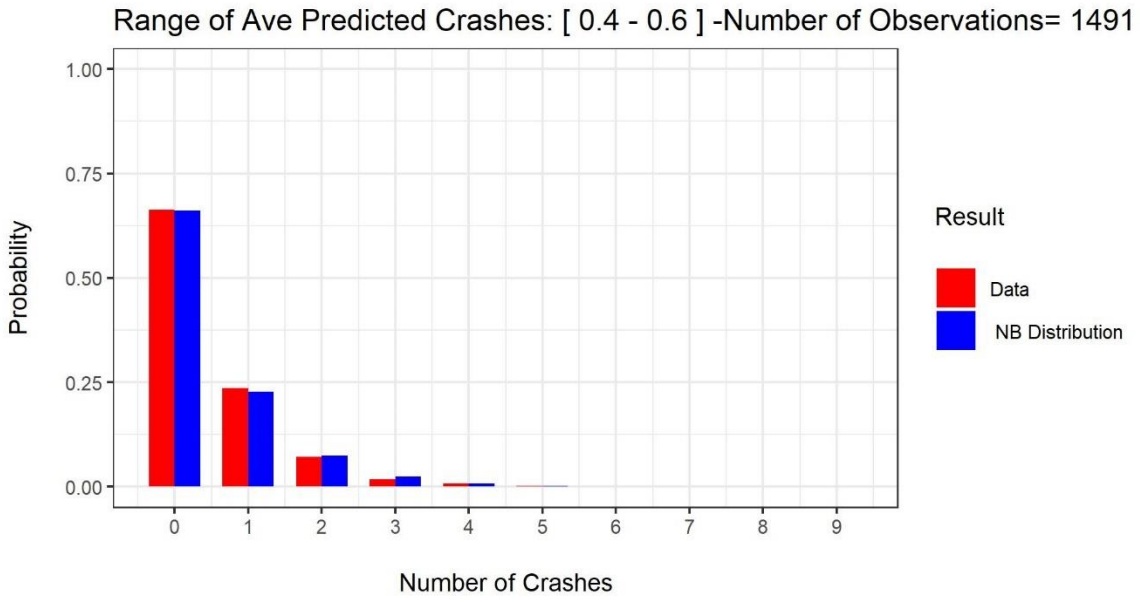


Figure 24 Comparison graph for the third range of average predicted crashes

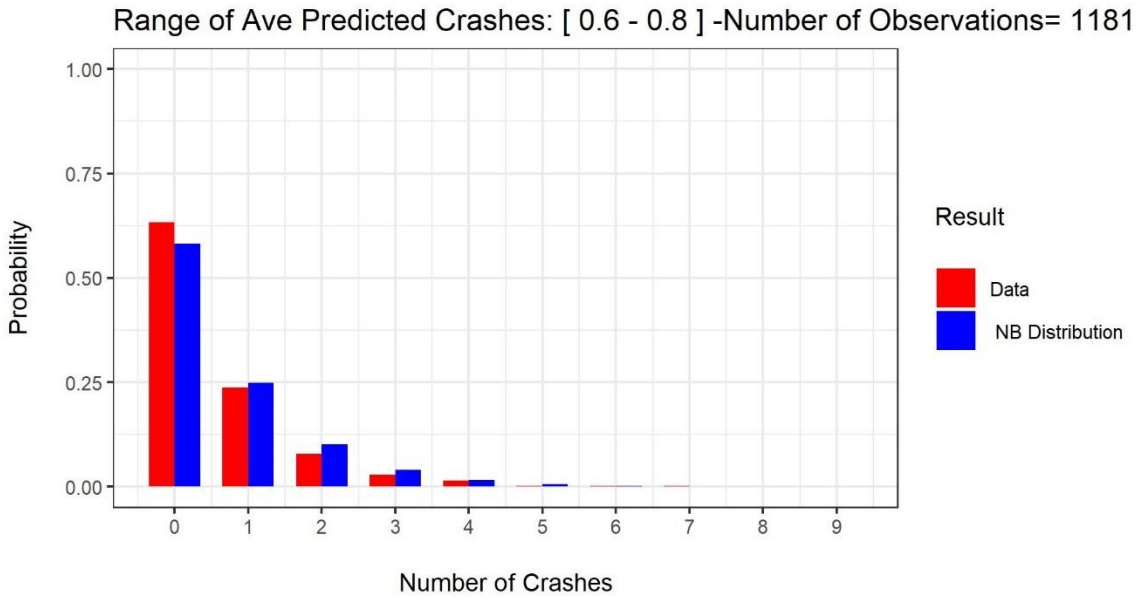


Figure 25 Comparison graph for the fourth range of average predicted crashes

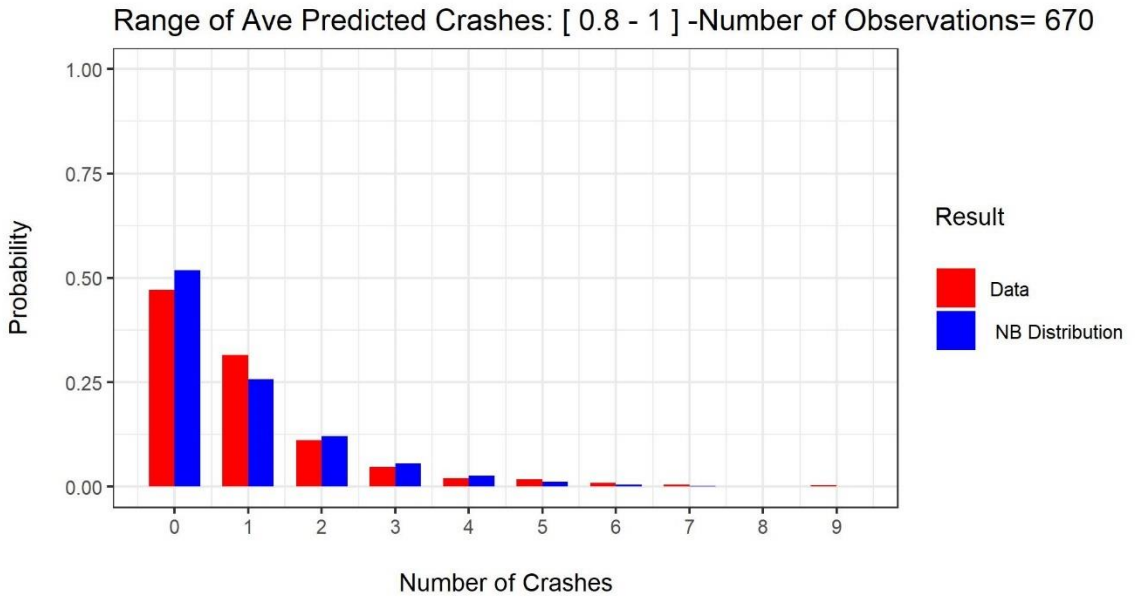


Figure 26 Comparison graph for the fifth range of average predicted crashes

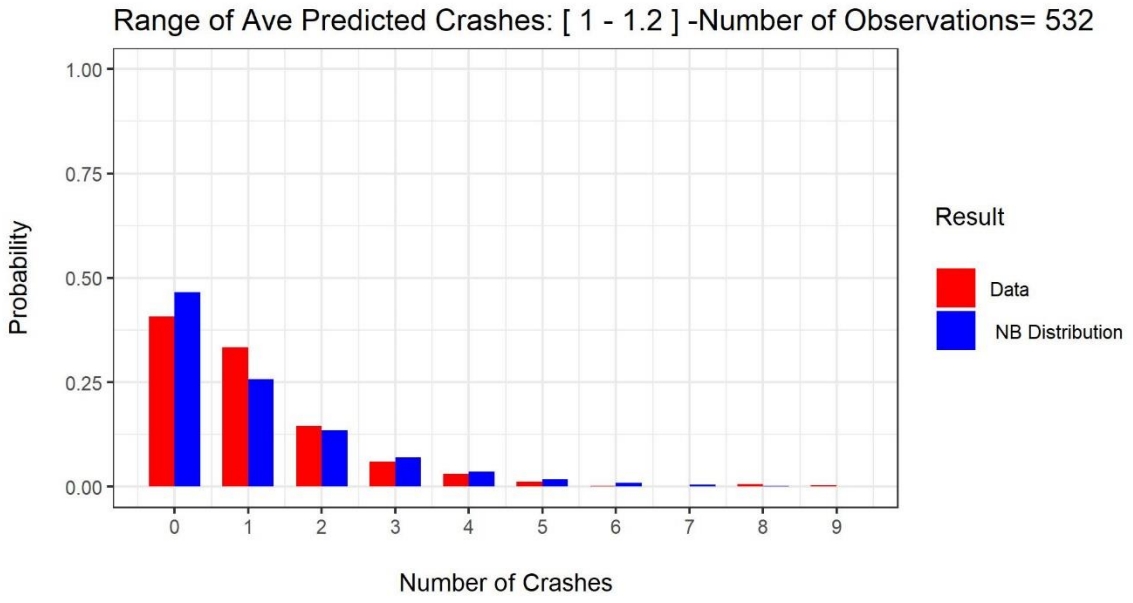


Figure 27 Comparison graph for the sixth range of average predicted crashes

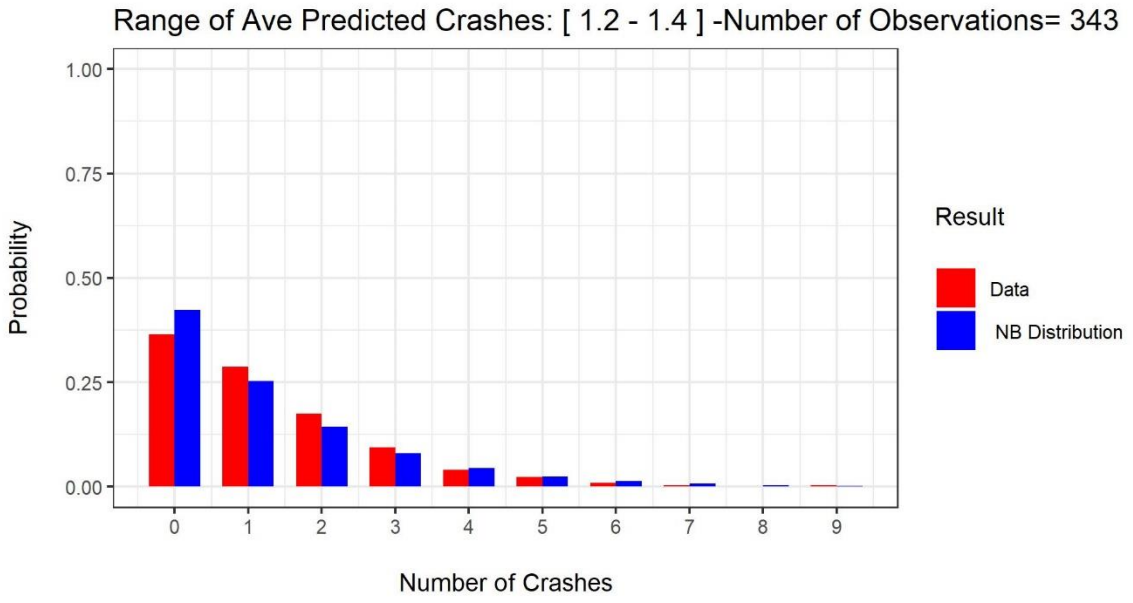


Figure 28 Comparison graph for the seventh range of average predicted crashes

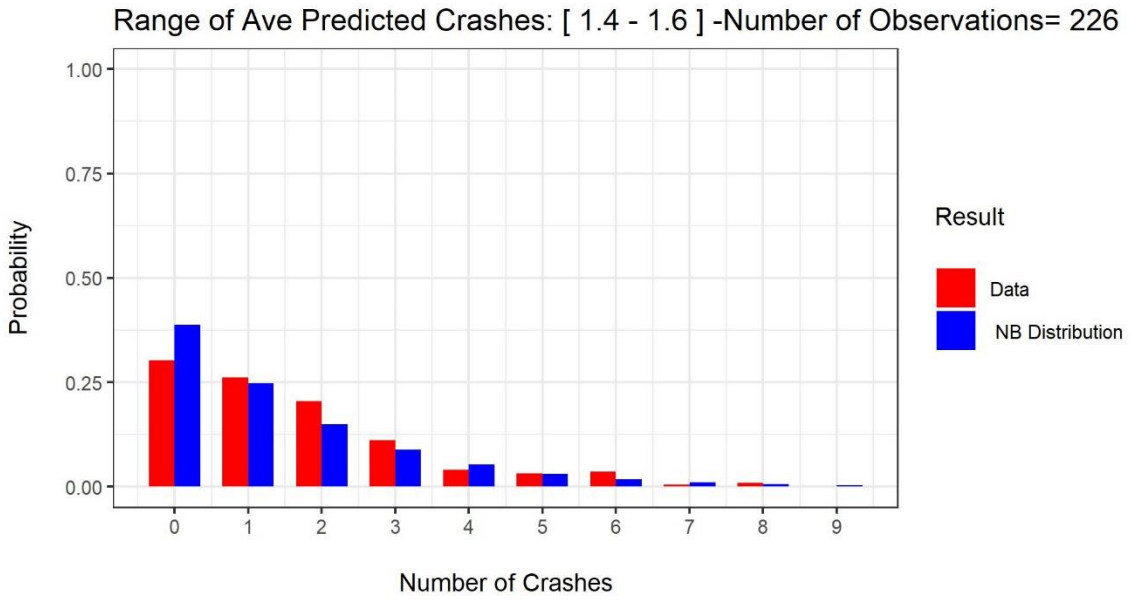


Figure 29 Comparison graph for the eighth range of average predicted crashes

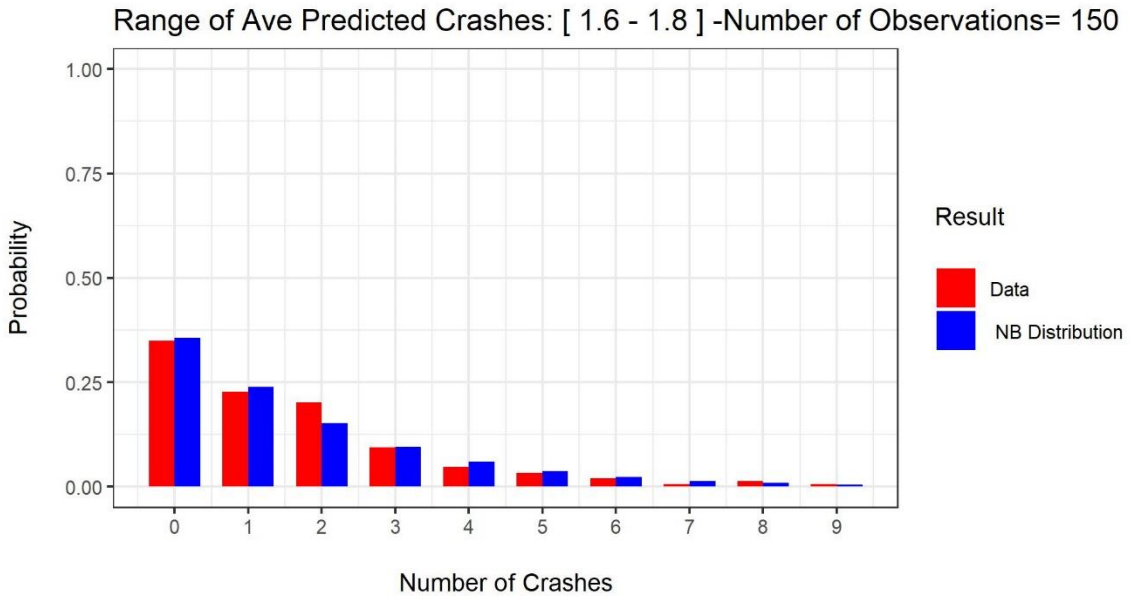


Figure 30 Comparison graph for the ninth range of average predicted crashes

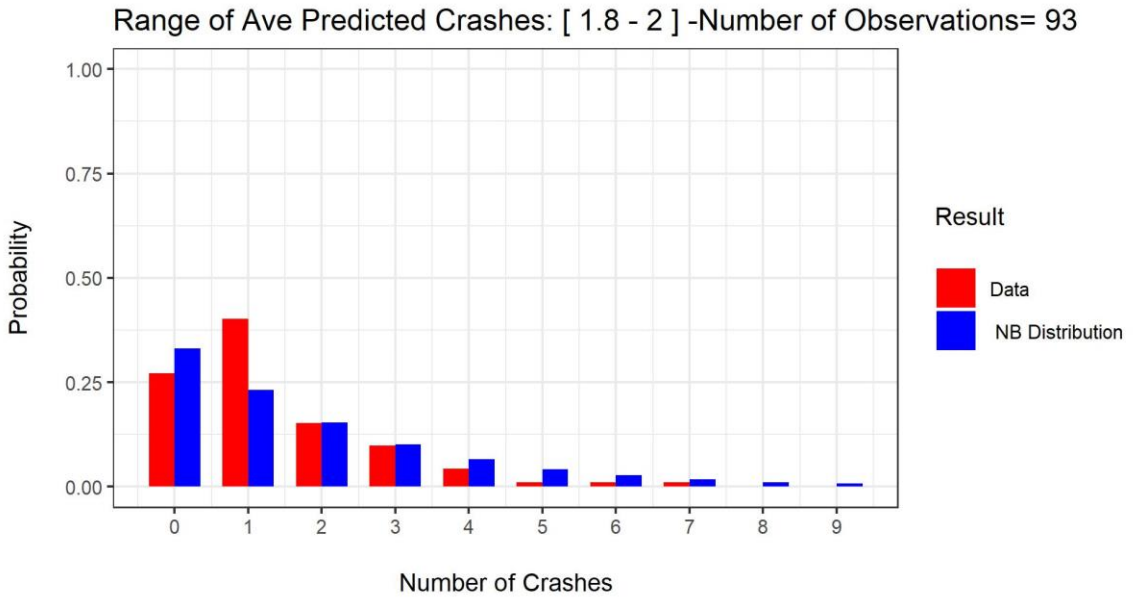


Figure 31 Comparison graph for the tenth range of average predicted crashes

Figure 32 shows a weighted average of the distributions presented for each range, where the number of observations provides the weights. As can be seen, the distributions are remarkably similar.

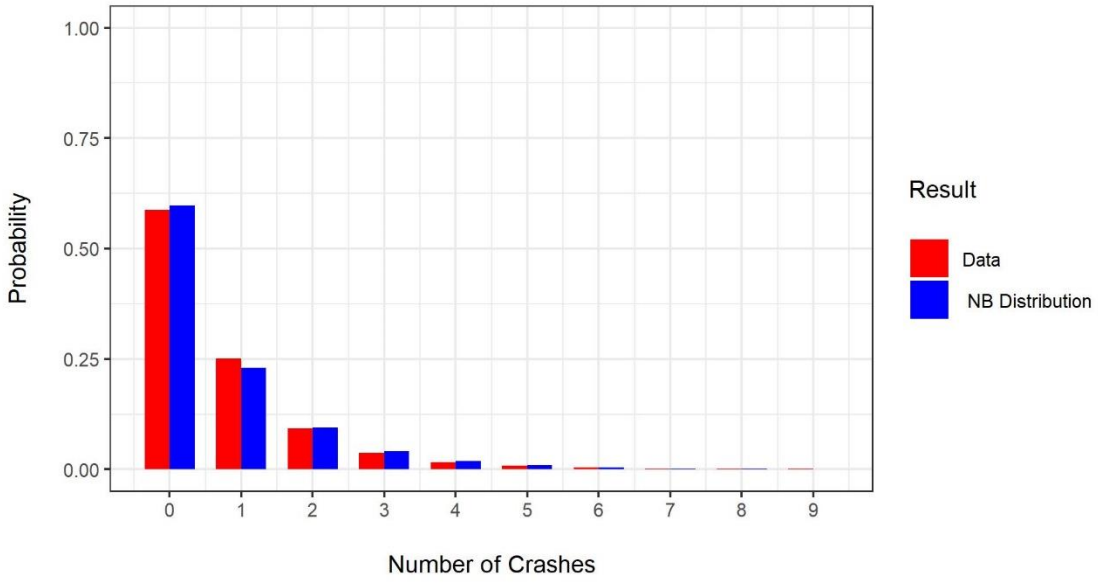


Figure 32 A weighted average of the distributions for all the ranges

Figure 33 presents a different way to look at the same information. In this figure, the average predicted and observed crash rates for each of the Figures 22-31 are computed using the definition of the expected value of a discrete probability distribution, that is $E[X] = \sum_i f_i X_i$, where f_i is the probability of observing X_i crashes. One advantage of this figure is that it is easy to compute an R^2 value. As shown in the figure, the $R^2 = 0.85$ which provides additional confirmation of the quality of the fitted model, even with the point with only 93 observations.

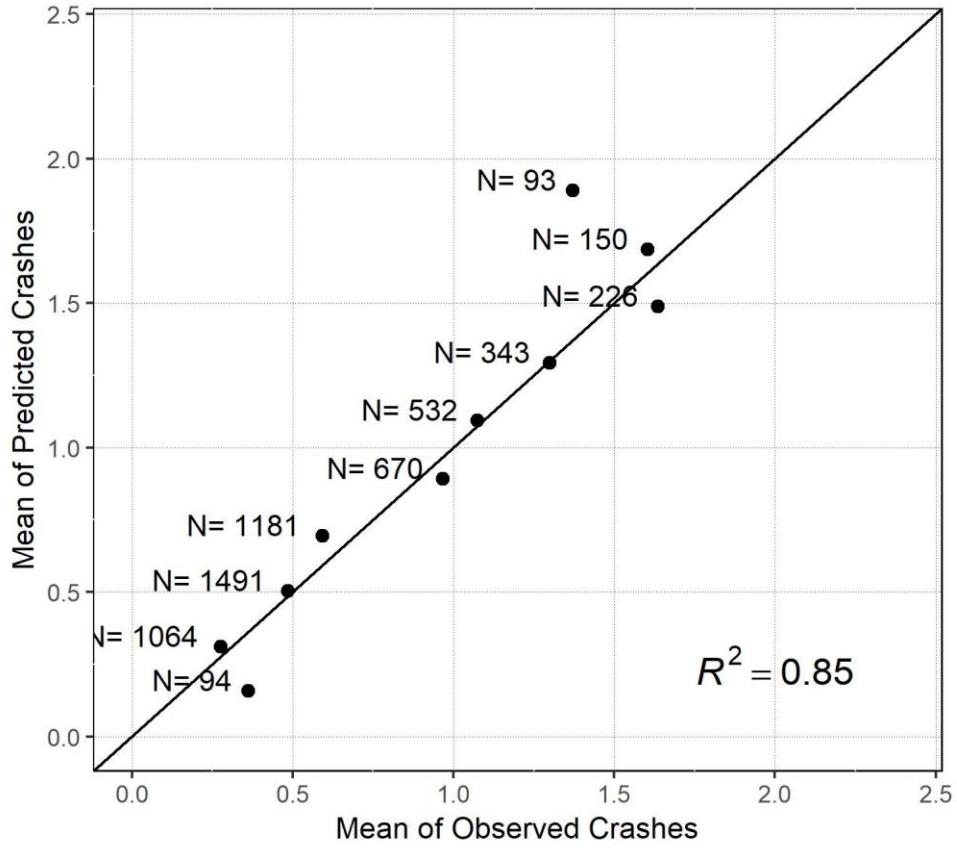


Figure 33 Mean of observed versus mean of predicted crashes

4.2.2 The zero-inflated negative binomial regression model

As mentioned earlier in section 2.1.3, the zero-inflated negative binomial regression model assumes that the probability of zero outcomes is the sum of the probabilities arising from two separate processes. In this case, the two processes are: 1) that a roadway segment is absolutely safe (the only possible outcome is zero) with a probability determined by a binary logit model, or 2) that no crashes are observed in the study period as a result of a zero-outcome arising from a process with negative binomial distribution.

The parameters of the zero-inflated regression model, estimated by maximum likelihood using the R-package pscl *(40)*, are presented in Figure 34. The results include two parts. The first part (count state) contains the negative binomial regression parameters for each of the variables along with standard errors, z-scores, and p-values for the parameters. The second part includes regression parameters, standard errors, z-scores, and p-values of the binary logit model for predicting the excess zeros in the zero-inflated model. The general form of the binary logit model is

$$P(X_i) = \frac{1}{1 + e^{-(U_i)}} \tag{25}$$

Where,

$$U_i = \beta_0 + \beta_1\ (Variable\ 1) + \beta_2\ (Variable\ 2) + \ \ldots\ \ + \beta_n\ (Variable\ N) \tag{26}$$

$P(X_i)$ in the probability of zero state condition for segment i, and U_i is the utility function for segment i. Equation 27 is obtained by substituting Equation 26 into Equation 25:

$$P(X_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1\ (Variable\ 1) + \beta_2 (Variable\ 2) + \ \dots\ + \beta_n (Variable\ N))}} \quad (27)$$

Equation 27 implies that an increase in the total value of utility function (i.e. U_i) will result in an increase in the probability of zero state condition for segment i. Therefore, the interpretation of results is based on the sign of the coefficients. These indicate whether their contribution to the utility function is positive or negative. For example, a unit increase in a variable with a negative coefficient decreases the utility function (i.e., U_i) and accordingly, it decreases the probability of being in a zero state condition (i.e., $P(X_i)$) for segment i.

4.2.2.1 Interpretation of results

This section presents the interpretation of results for both parts of the model (i.e., the count sub-model and the zero-state binary logit sub-model). It explores the factors that affect the frequency of RwD crashes in a count model as well as the factors that influence the zero-state condition.

4.2.2.1.1 Zero-inflated model - count sub-model

The count sub-model (i.e., the top block in Figure 34) identifies the statistically significant variables and their estimated coefficients. This part of the zero-inflated negative binomial model is almost identical, in terms of the variables included and the signs of the parameters, to the negative binomial regression model presented before in Section 4.2.1. The only difference between the count sub-model in the zero-inflated model and the negative binomial model is that the lane width is not significant in the count sub-model of the zero-inflated model. Instead, the lane width appears in the zero state sub-model.

Because the interpretations of results are identical, this section skips to the second part of the model to avoid repetition.

```
Count model coefficients (negbin with log link):
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -5.0093809  0.3466729 -14.450  < 2e-16 ***
log(AADT)           -0.5360362  0.0359227 -14.922  < 2e-16 ***
Abs_Curvature        0.0976840  0.0101090   9.663  < 2e-16 ***
I(Abs_Curvature^2)  -0.0027824  0.0003384  -8.222  < 2e-16 ***
PercentageTruck      0.0107546  0.0040322   2.667  0.00765 **
Grade               -0.0411173  0.0067845  -6.060 1.36e-09 ***
I(Grade^2)           0.0041577  0.0016363   2.541  0.01105 *
PaintedMedianWidth  -0.0449389  0.0209129  -2.149  0.03165 *
Shoulder_width      -0.0578050  0.0087761  -6.587 4.50e-11 ***
Curvature_ENV_SD     0.0926397  0.0229602   4.035 5.46e-05 ***
Curvature_ENV_Ave   -0.0921279  0.0289148  -3.186  0.00144 **
IRI_ENV_SD          -0.0038985  0.0012752  -3.057  0.00223 **
IRI_ENV_Ave          0.0017575  0.0006357   2.765  0.00570 **
Mean_FrictionDemand  2.4694700  0.5681998   4.346 1.39e-05 ***
PofLengthOfGuardrail -0.1472005  0.0625591  -2.353  0.01862 *
PofLengthOfMedian   -0.4154423  0.1005698  -4.131 3.61e-05 ***
Log(theta)           0.3391572  0.0649708   5.220 1.79e-07 ***

Zero-inflation model coefficients (binomial with logit link):
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        8.20595    2.22452   3.689 0.000225 ***
log(AADT)         -3.57741    0.31550 -11.339  < 2e-16 ***
Abs_Curvature     -0.05103    0.01941  -2.629 0.008560 **
PercentageTruck   -0.12127    0.05519  -2.197 0.028000 *
LaneWidth          0.95924    0.16981   5.649 1.62e-08 ***
Shoulder_width    -0.29198    0.08170  -3.574 0.000352 ***
Curvature_ENV_Ave  0.11063    0.02711   4.080 4.50e-05 ***
Grade_ENV_SD       0.39467    0.10804   3.653 0.000259 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 1.4038
Number of iterations in BFGS optimization: 49
Log-likelihood: -6759 on 25 Df
```

Figure 34 Estimation results for the zero-inflated negative binomial model

4.2.2.1.2 Zero-inflated model – binary logit sub-model

This section describes the parameters corresponding to the zero-crash state in the second part. The sign of the estimated coefficient for the logarithm of AADT is negative. So, the log odds of being an excessive zero would decrease approximately by 3.57 for every unit increase in the log of AADT. In other words, a segment with a higher volume of traffic is associated with a smaller chance that the zero-observed crash is due to the inherent safety of the segment. The same interpretation is also valid for the other three variables with negative signs: Abs_Curvature (-0.051), PercentageTrucks (-0.121), and Shoulder_Width (-0.292). These results mean that for a segment with a higher percentage of trucks, a higher absolute value of curvature, or a larger shoulder width, the zero-crash observation is associated more with the failure to observe a crash during the study period rather than the inherent safety of the segment. The effect of shoulder width deserves further scrutiny. The model indicates that the log odds of being an excessive zero would decrease approximately by 0.29 for every unit increase in the shoulder width. The interesting point is that a wider shoulder width has already been found to reduce the frequency of RwD crashes in the count model; while, it has been found to reduce the odds of being inherently safe in the zero-inflation model. A possible explanation to justify this finding is that wider shoulder width provides more confidence for drivers to drive at higher speeds, and accordingly, it increases the chances of drivers to be involved in a RwD crash.

On the other hand, a unit increase in lane width, the standard deviation of grade (environmental variable), or the absolute value of curvature (environmental variable) would increase the log odds of being a zero-crash state approximately by 0.95, 0.39 and

0.11 respectively. Therefore, a segment with wider lanes or located after a curvy road is inherently safer. In addition, the model suggests a higher standard deviation of grades before each segment can increase the probability that the zero-crash observation is due to the segment's inherent safety. A higher standard deviation of grade may result from changes in the signs of the previous segments' grade. Therefore, a possible explanation is that roads with many ups and downs make drivers feel uncomfortable in part because of lower available sight distances, so they tend to pay more attention to the road's features.

4.2.2.2 Vuong statistic

The negative binomial regression model and the zero-inflated negative binomial regression model are not nested. To evaluate the appropriateness of using zero-inflated model over the other, Vuong suggests a test statistic for non-nested models that is appropriate for this setting where the distributions of models are specified (e.g., Poisson or negative binomial) *(20)*. A value m_i is computed for each observation using Equation 28.

$$m_i = ln\left(\frac{f_1(y_i|X_i)}{f_2(y_i|X_i)}\right) \tag{28}$$

where, $f_1(y_i|X_i)$ is the probability density function of the negative binomial model and $f_2(y_i|X_i)$ is the probability density function of the count sub-model of the zero-inflated model. The Vuong statistic is then calculated using Equation 29.

$$V = \frac{\sqrt{n}\left[\left(\frac{1}{n}\right)\sum_{i=1}^{n} m_i\right]}{\sqrt{\left(\frac{1}{n}\right)\sum_{i=1}^{n}(m_i - \overline{m})^2}} = \frac{\sqrt{n}(\overline{m})}{S_m} \tag{29}$$

where, $\overline{m}$ is the mean:

$$\bar{m} = \left(\frac{1}{n}\right)\sum_{i=1}^{n} m_i \quad (30)$$

and n is the number of segments.

Shankar et al. provide a model selection guideline based on the possible values of the Vuong test and the t-statistics of the overdispersion parameter. This test is suitable for situations where the absolute value of the t-statistic of the negative binomial's overdispersion parameter is higher than 1.96 (the 95 percent confidence level). The test supports a zero-inflated model if the V is greater than 1.96 and supports the negative binomial if it is smaller than -1.96. The test is inconclusive for V values in between.

In this case, the Vuong statistic value is 5.43, and the t-statistic of the negative binomial's overdispersion parameter is higher than 1.96, thus supporting the appropriateness of zero-inflated model over the negative binomial regression for the 0.2-mile segment length.

4.2.2.3 Model evaluation

An evaluation approach for the negative binomial model was introduced in Section 4.2.1.2. The same approach is used to evaluate the results of the zero-inflated negative binomial model. Figures 35-44 exhibit acceptable similarities between the two probability distributions especially for the ranges with a higher number of observations. The main difference between these graphs with those for the negative binomial model is that a higher proportion of segments are predicted with zero crashes.

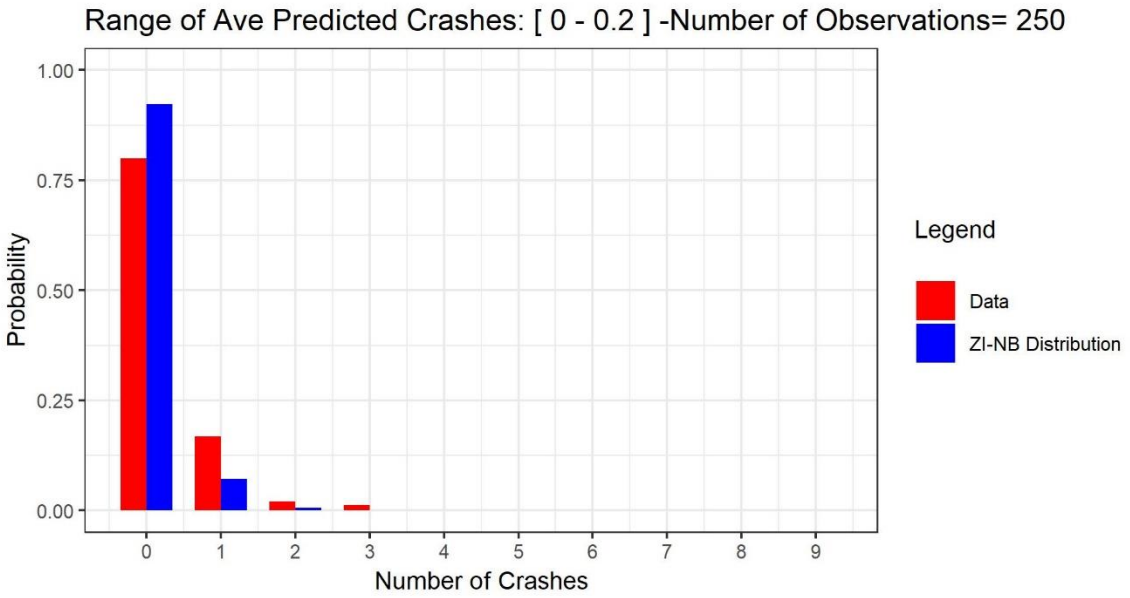


Figure 35 Comparison graph for the first range of average predicted crashes

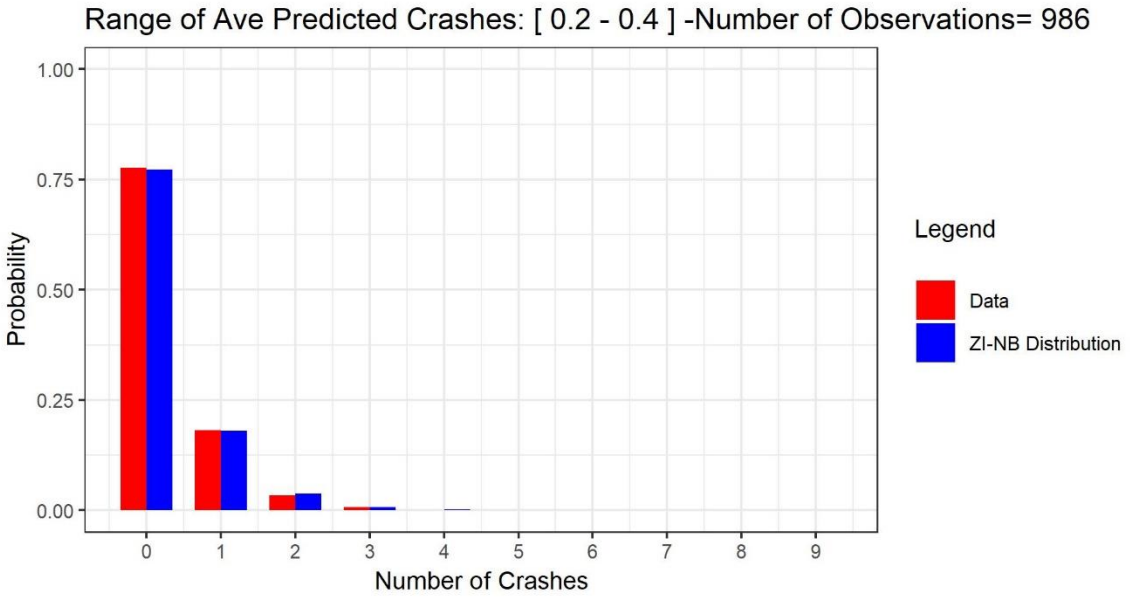


Figure 36 Comparison graph for the second range of average predicted crashes

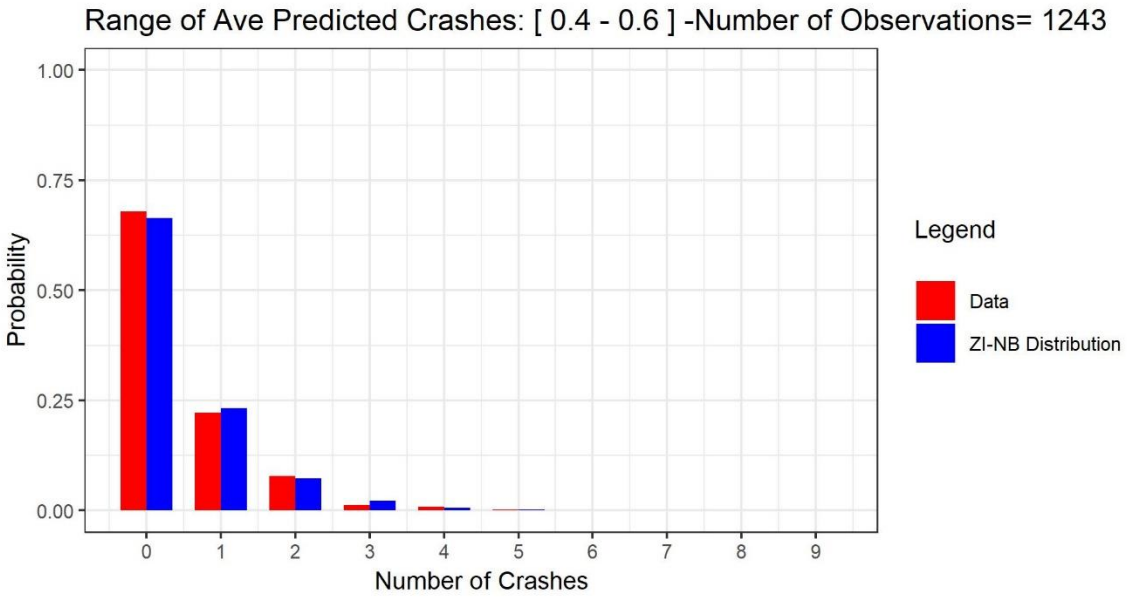


Figure 37 Comparison graph for the third range of average predicted crashes

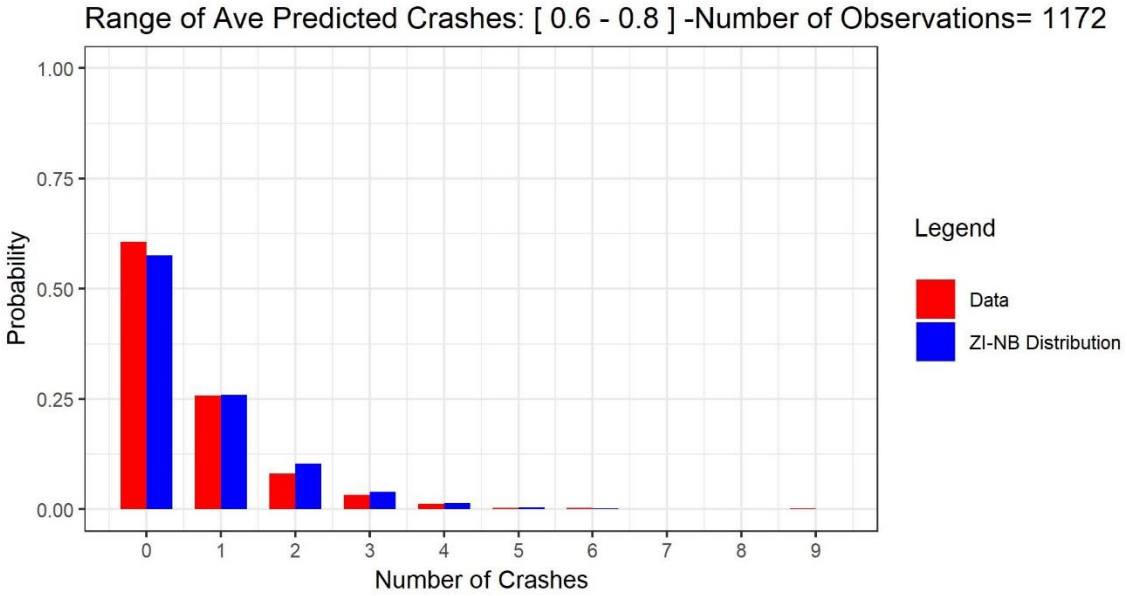


Figure 38 Comparison graph for the forth range of average predicted crashes

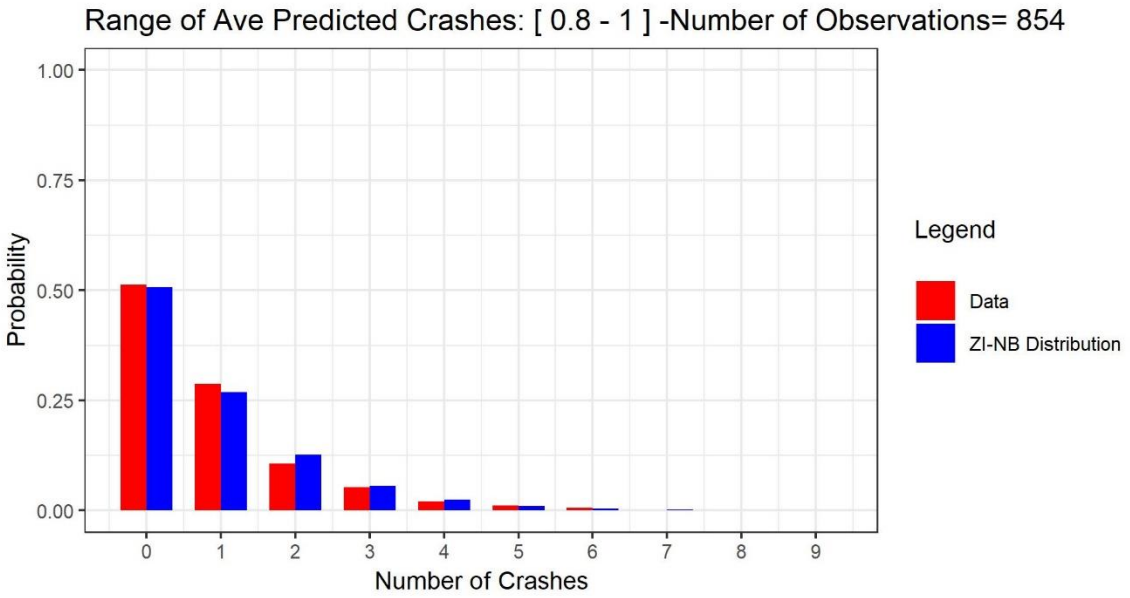


Figure 39 Comparison graph for the fifth range of average predicted crashes

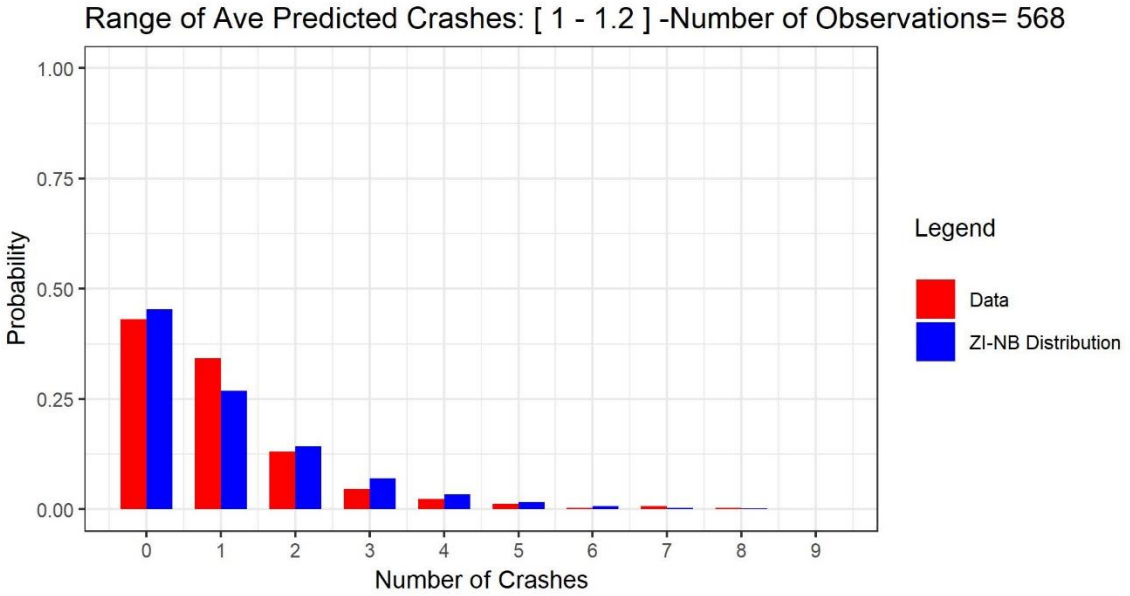


Figure 40 Comparison graph for the sixth range of average predicted crashes

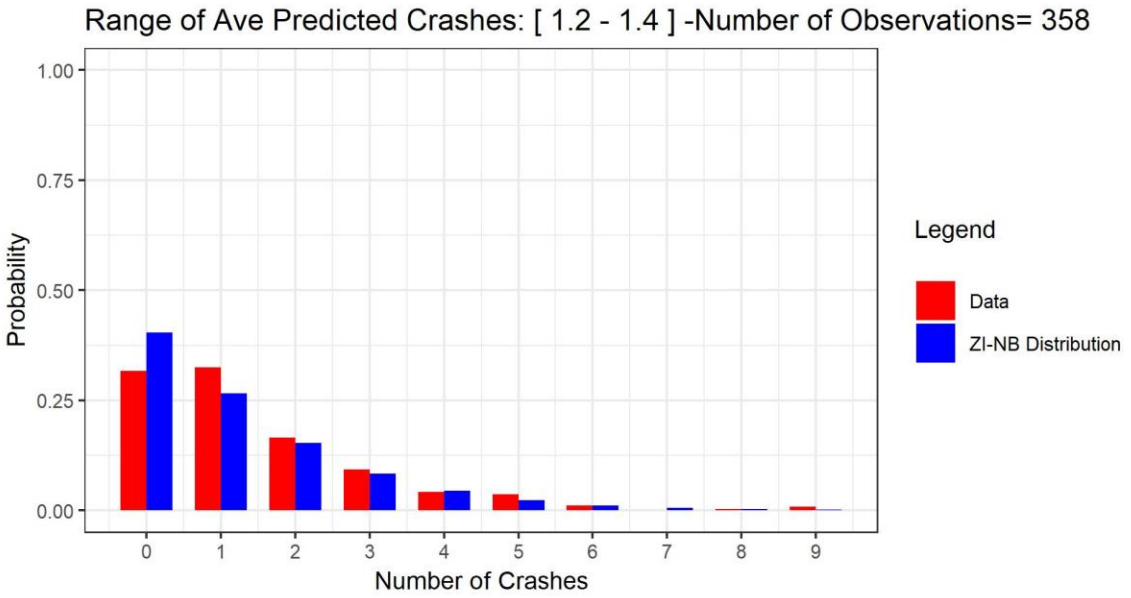


Figure 41 Comparison graph for the seventh range of average predicted crashes

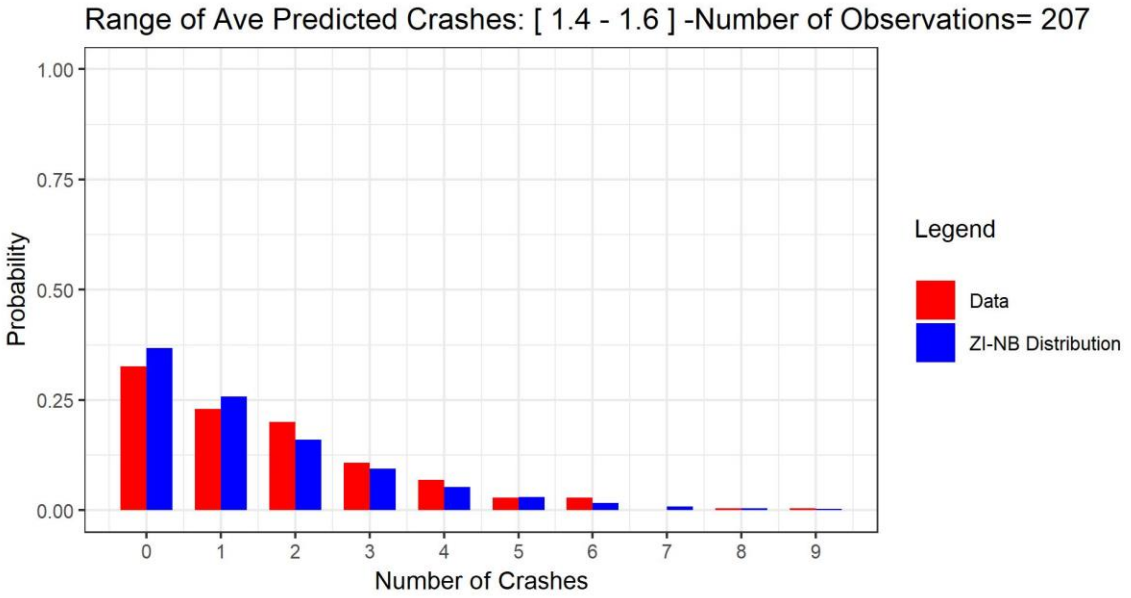


Figure 42 Comparison graph for the eighth range of average predicted crashes

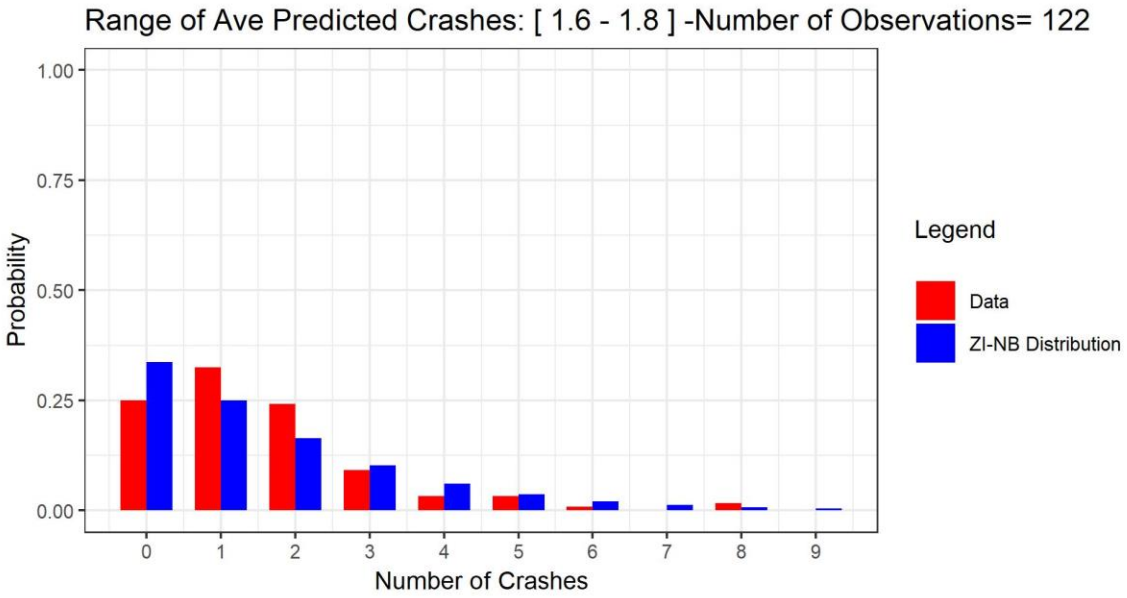


Figure 43 Comparison graph for the ninth range of average predicted crashes

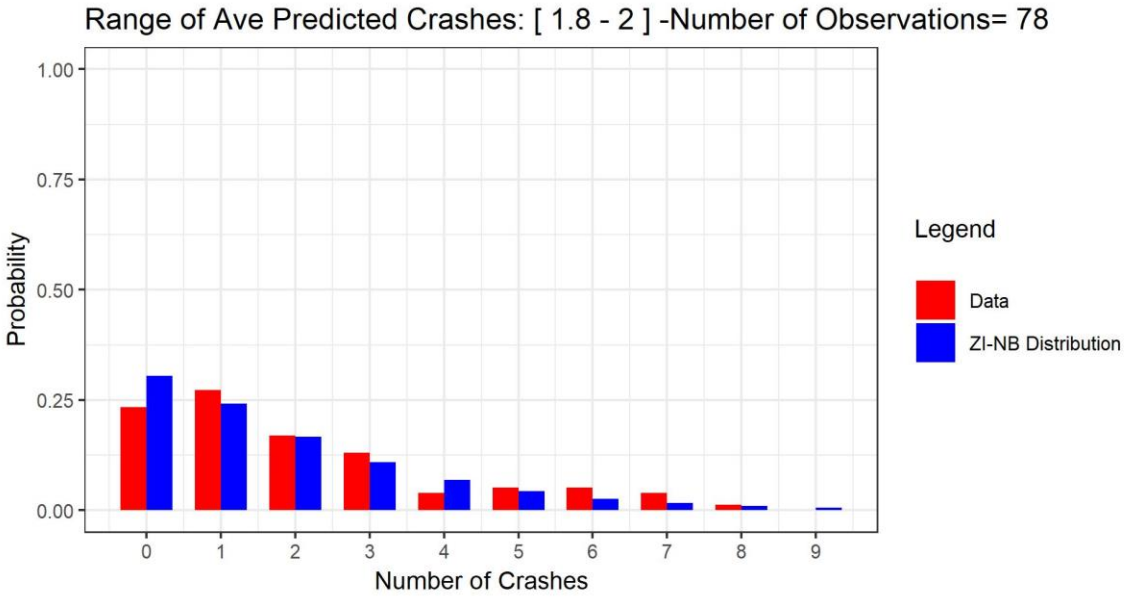


Figure 44 Comparison graph for the Tenth range of average predicted crashes

Figure 45 illustrates a weighted average of the distributions presented for each range, where the number of observations provides the weights. As can be seen, the distributions are extremely similar.

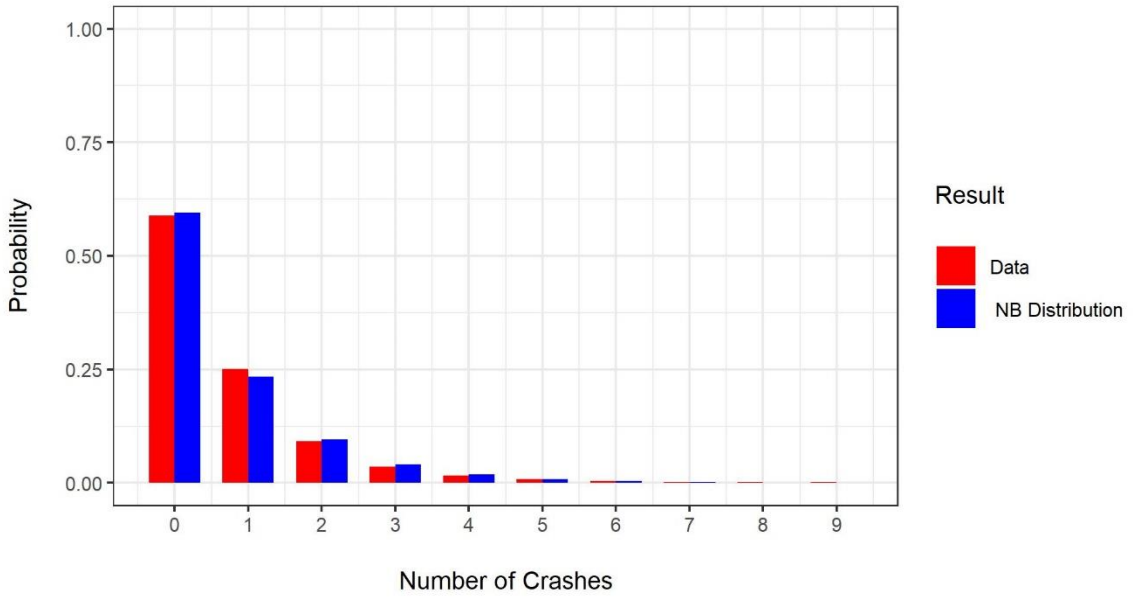


Figure 45 A weighted average of the distributions for all the ranges

In Figure 46, the average predicted and observed crash rates for each of the Figures 35-44 is computed. As shown in the figure, the $R^2 = 0.96$ which provides additional confirmation of the quality of the zero-inflated model, and its superiority in terms of the statistics to the negative binomial regression model with $R^2 = 0.85$.

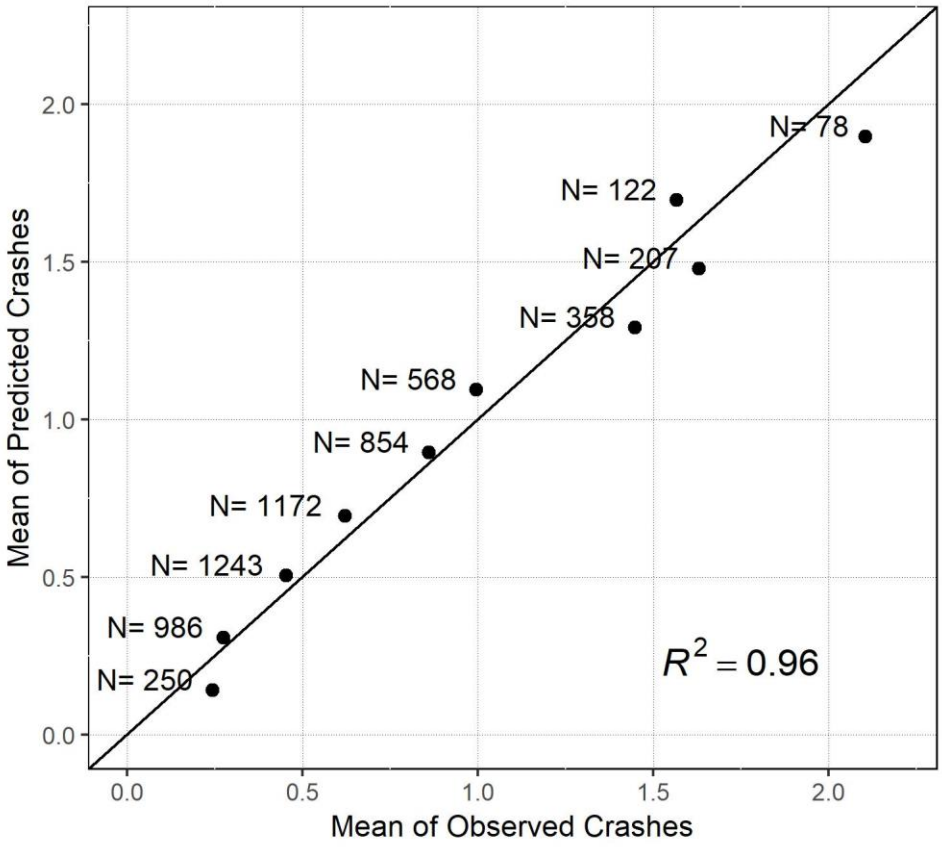


Figure 46 Observed versus predicted graph

4.2.3 The mixed-effects negative binomial regression model

The parameters of a mixed-effects negative binomial regression model were estimated using the Laplace approximation method with the R-package lme4 *(41)*. Laplace approximation is a mathematical method for integral approximation *(42)*. The estimation results are presented in Figure 47. The estimation of random effects including the variance and standard deviation of random parameters is presented in the first part of Figure 47. The second part presents the fix-effects parameters estimation, their corresponding standard errors, z values, and p-values.

```
     AIC      BIC   logLik deviance df.resid
 13483.4  13610.6  -6722.7  13445.4     5945

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.0205 -0.5686 -0.4239  0.5007  3.8398

Random effects:
 Groups      Name                Variance  Std.Dev.
 numbering   LaneWidth            0.006122 0.07824
 numbering.1 Mean_FrictionDemand 22.140762 4.70540
Number of obs: 5964, groups:  numbering, 5964

Fixed effects:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)               -5.7304688  0.3670145 -15.614  < 2e-16 ***
log(AADT)                 -0.3405685  0.0307027 -11.092  < 2e-16 ***
Abs_Curvature              0.0786790  0.0102760   7.657 1.91e-14 ***
I(Abs_Curvature^2)        -0.0017412  0.0003076  -5.661 1.50e-08 ***
PercentageTruck            0.0172801  0.0041434   4.171 3.04e-05 ***
Grade                     -0.0342758  0.0068678  -4.991 6.01e-07 ***
LaneWidth                 -0.1239204  0.0252327  -4.911 9.06e-07 ***
Shoulder_width            -0.0354807  0.0087610  -4.050 5.12e-05 ***
Curvature_ENV_SD           0.0631330  0.0215967   2.923  0.00346 **
Curvature_ENV_Ave         -0.1087721  0.0257081  -4.231 2.33e-05 ***
IRI_ENV_Ave                0.0012302  0.0004316   2.851  0.00436 **
Mean_FrictionDemand        2.1898217  0.8212668   2.666  0.00767 **
PofLengthOfGuardrail      -0.1832131  0.0668342  -2.741  0.00612 **
PofLengthOfMedian         -0.4721022  0.1080126  -4.371 1.24e-05 ***
PofLengthOfACShoulderType -0.1875690  0.0616842  -3.041  0.00236 **
PaintedMedianWidth        -0.0447642  0.0223571  -2.002  0.04526 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 47 The estimation results for the mixed-effects negative binomial model

In the model estimation process, a random-effect was initially assumed for each parameter, and then all the parameters were estimated. In the second step, the likelihood ratio test was used to determine the significance of each random term in the model. After performing several likelihood ratio tests, only those random-effects that passed the likelihood ratio tests remained in the final model. In other words, the random-effects were considered only for parameters that their estimated standard deviations were statistically

significant at a 5 percent significance level. More detailed interpretations of the results are provided in the next section.

4.2.3.1 Interpretation of results

This section presents the interpretation of results for both parts of the model (i.e., the fixed-effects and the random-effects).

4.2.3.1.1 Fixed-Effects

As mentioned earlier, the fixed-effects are presented in the last part of the model's output. The interpretation of statistically significant parameter estimates and their signs are similar to the result of the negative binomial regression model presented before in Section 4.2.1.1. The main difference between the fixed-effects in the mixed-effects model and the negative binomial model is that the proportion of the total length of sections with painted medians is not statistically significant in the mixed-effects model. Because the interpretations of results are almost identical to what was presented before in Section 4.2.1.1, this section discusses only the random-effects interpretation to avoid repetition.

4.2.3.1.2 Random-effects

As mentioned earlier, the mixed-effects formulation allows the modeling of the parameter estimates to vary across the population of roadway segments by modeling each parameter as the sum of a fixed effect plus a random effect with a certain probability distribution (usually assumed normal). This formulation is intended to capture the unobserved heterogeneities across the segments. The estimated random-effects parameters are presented in the first part of Figure 47.

The results show that the random effects are statistically significant for two explanatory variables: lane width and side friction demand. Note that the estimated standard deviation of the random effect for side friction demand (4.7054) is quite large relative to the estimated value of the corresponding fixed effect (2.1898). On the other hand, for the lane width parameter estimate, the standard deviation of the random effect (0.0782) is smaller than the corresponding fixed effect (0.1239). This means that in both cases, there is a non-negligible probability of the sign of the estimated parameter changing across the population of segments.

The estimated normal Probability Density Function (PDF) of the lane width parameter is depicted in Figure 48. The figure shows that according to the estimation results there is a 0.943 probability that the parameter is negative[1]. Accordingly, the area under the PDF for values greater than 0 is equal to:

$$P\ (the\ coefficient\ of\ lane\ width\ >\ 0)\ = 1 -\ (the\ probability\ of\ grey\ area)$$

$$=\ 1 - 0.943 = 0.057$$

This result confirms that, in general, wider lanes help to reduce the rate of RwD crashes. However, the uncertainty is high enough that in a small proportion of cases (5.7 percent) a slight increase in the frequency of crashes is predicted with wider lanes. A possible explanation is that in a few instances the positive effects of a narrow traffic lane (e.g., persuading

[1] The probability that a continuous random variable falls in the interval between a and b is equal to the area under the pdf curve between a and b

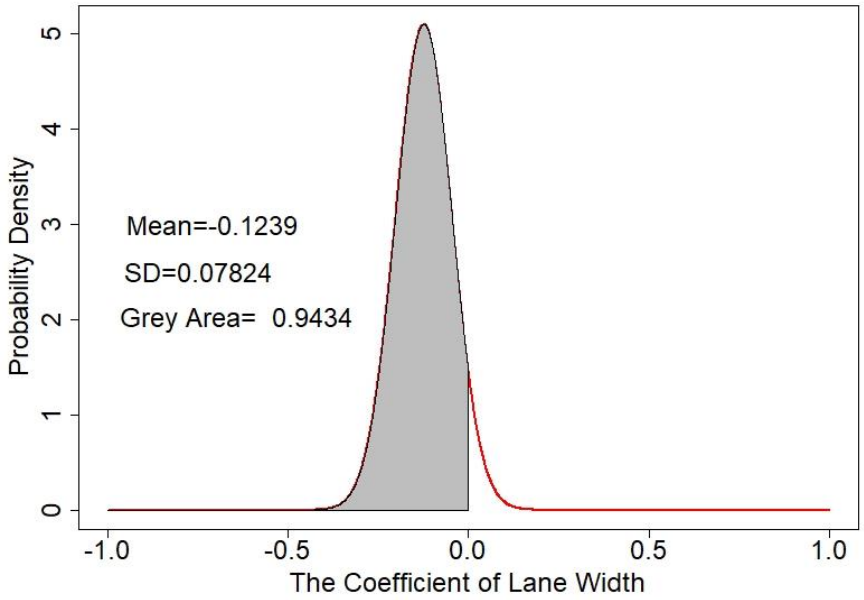


Figure 48 The normal distribution of lane width' coefficient

drivers to drive more attentively) can be more than its adverse effects (e.g., a higher rate of conflicts between both directions of travels).

The side friction demand resulted in a random parameter that is normally distributed, with a mean 2.19 and standard deviation of 4.22. The PDF of mean side friction demand is depicted in Figure 49. Given these distributional parameters, 32.08 percent of the distribution is less than 0 and 67.92 percent is greater than 0. This implies that for most of the road segments an increase in side friction demand increases the rate of RwD crashes. However, that leaves almost a third of cases for which the model predicts a decrease in the rate of RwD crashes with an increase on side friction demand. A possible explanation for the higher uncertainty in the effect of side friction demand may be related to two issues:

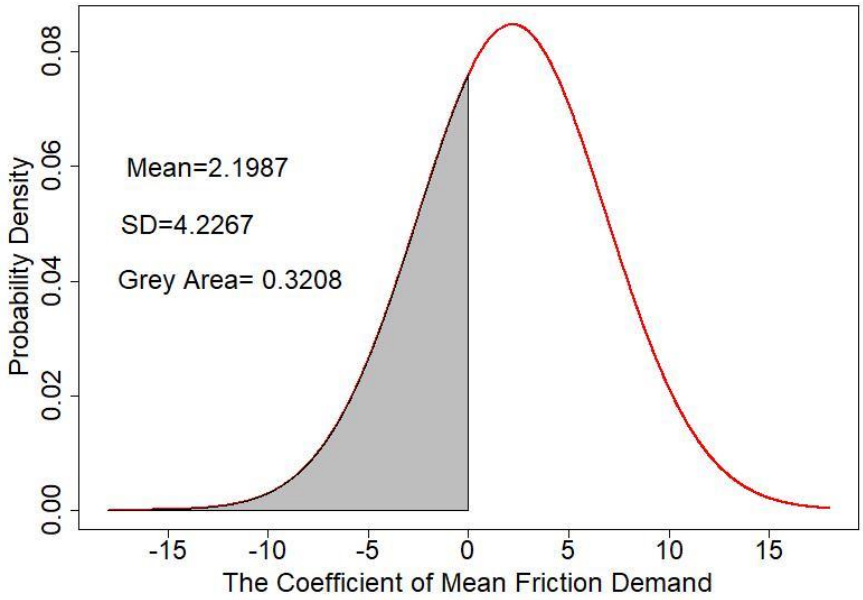


Figure 49 The normal distribution of friction demand's coefficient

First, the side friction demand incorporates in part the effect of curvature (which is already in the model). Although the two variables are not highly correlated, they both incorporate a common factor. Second, the side friction demand also depends on other information such as speed and superelevation that has some limitations. Remember that the speed of the van used for collecting pavement distresses was used as a surrogate of a representative speed on curves. In some cases, this estimation may not be very accurate.

4.2.3.2 Model evaluation

An evaluation approach for the negative binomial model was introduced in Section 4.2.1.2. The same approach is used to evaluate the results of the mixed-effects negative binomial model. In this case, the main difference is that only the fixed-effects are used for the prediction because the random effects are random, and they cannot be used for

prediction. Figures 50-59 show the results for a segmentation of the data based on predicted values (using only the fixed effects) in ranges of 0.2 crashes. As seen in these figures, the predicted and observed distributions for each data segment show more dissimilarities between the two probability distributions in comparison with the previous two models. This is common with mixed-effects models as the better statistics are obtained by modeling the distribution of the unobserved heterogeneity in some parameters at the expense of the fit obtained with just the fixed-effects. Note that although the modeling does not require the computation of the random-effects for each roadway segment, the estimation approach is as if a different random effect was estimated for each segment. Thus, estimation with just the fixed-effects typically results in a poorer fit. This is compounded by the fact that the estimation of the random-effects distributional parameters is typically done at the expense of the loss of statistical significance of some fixed-effects (as was the case in this research). It is important to note, however, that the lower fit does not mean the mixed-effects model is inferior to the others. In fact, its statistics (e.g., the log-likelihood) are better than for the other two models. It must be remembered that although the model fit is important so is obtaining a model that captures the data generation process accurately, including the level of uncertainty. Thus, model selection becomes a difficult task with no clear answer.

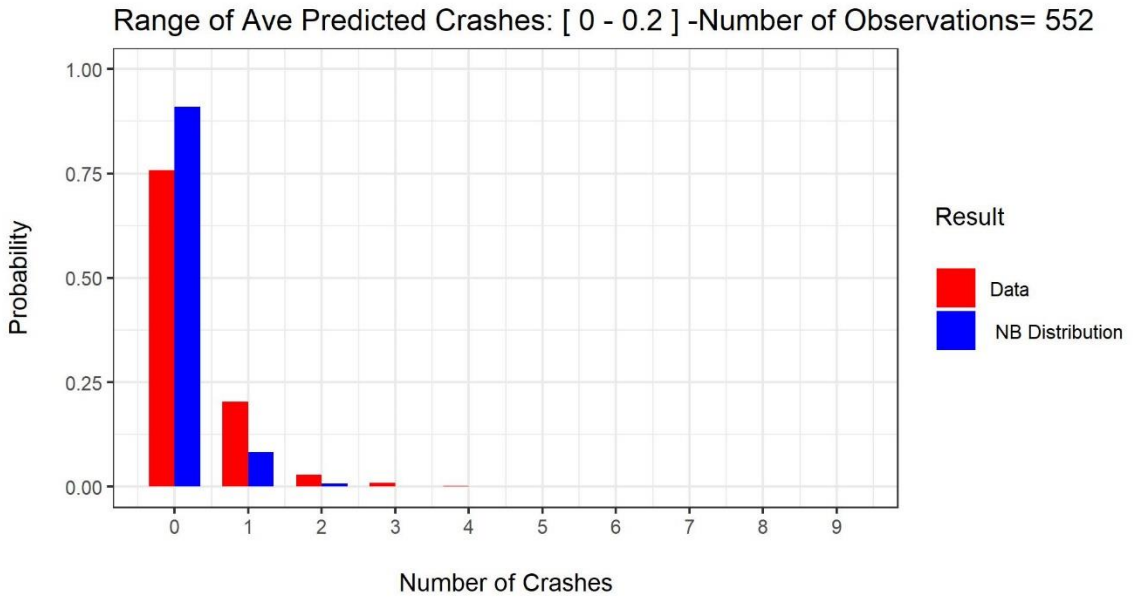


Figure 50 Comparison graph for the first range of average predicted crashes

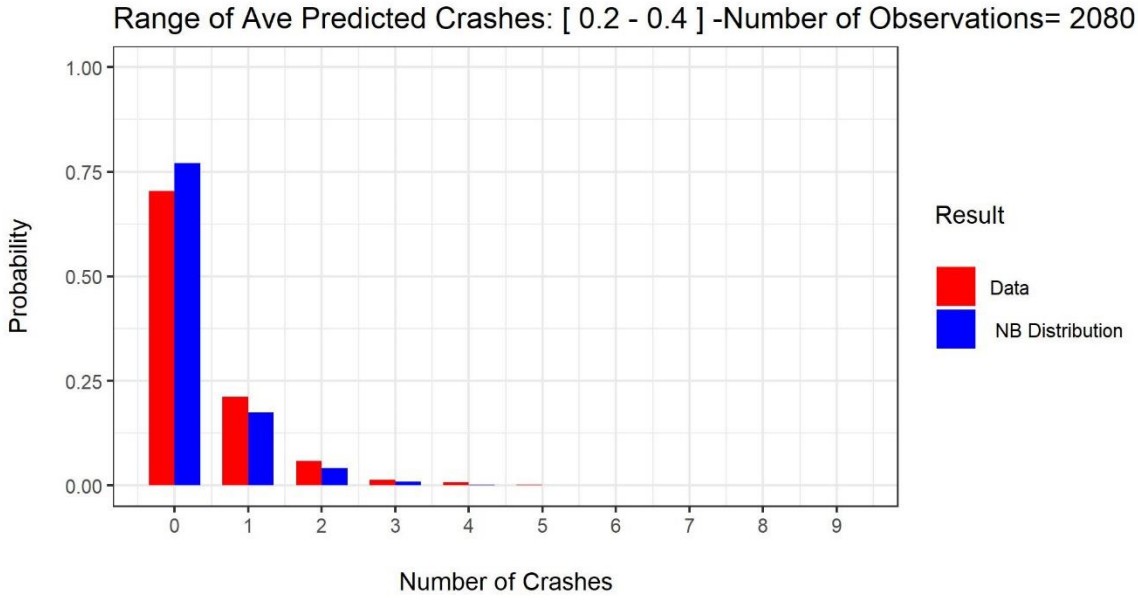


Figure 51 Comparison graph for the second range of average predicted crashes

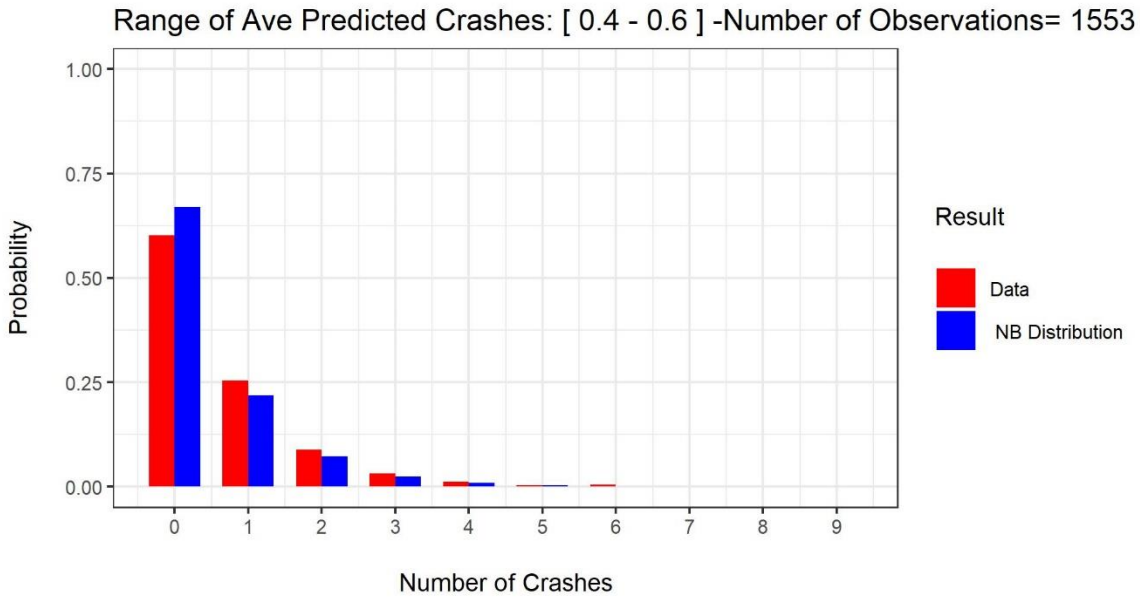


Figure 52 Comparison graph for the third range of average predicted crashes

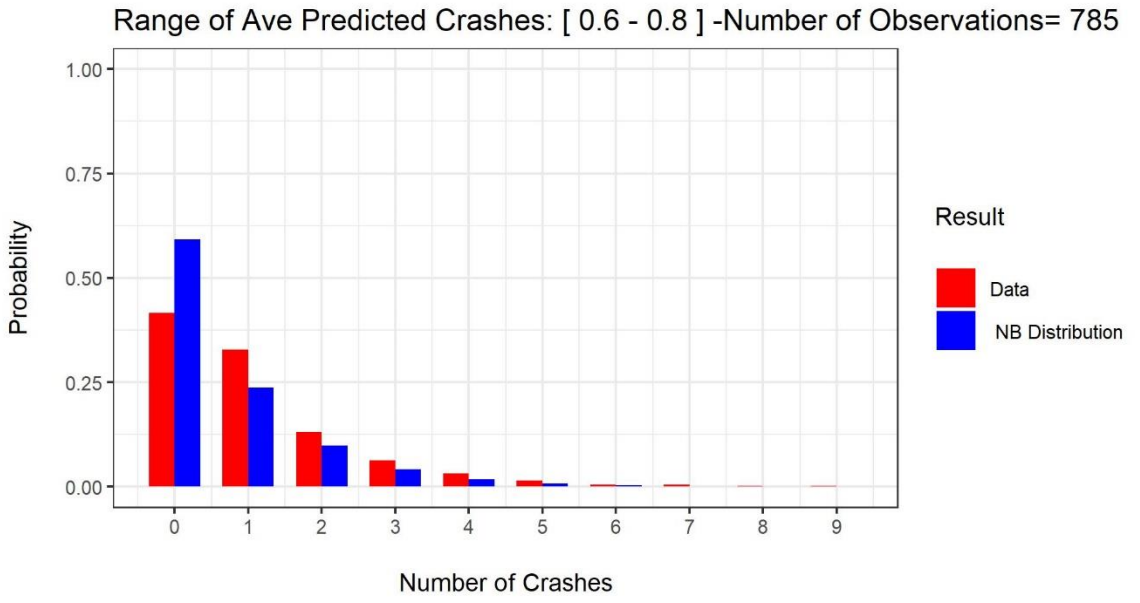


Figure 53 Comparison graph for the fourth range of average predicted crashes

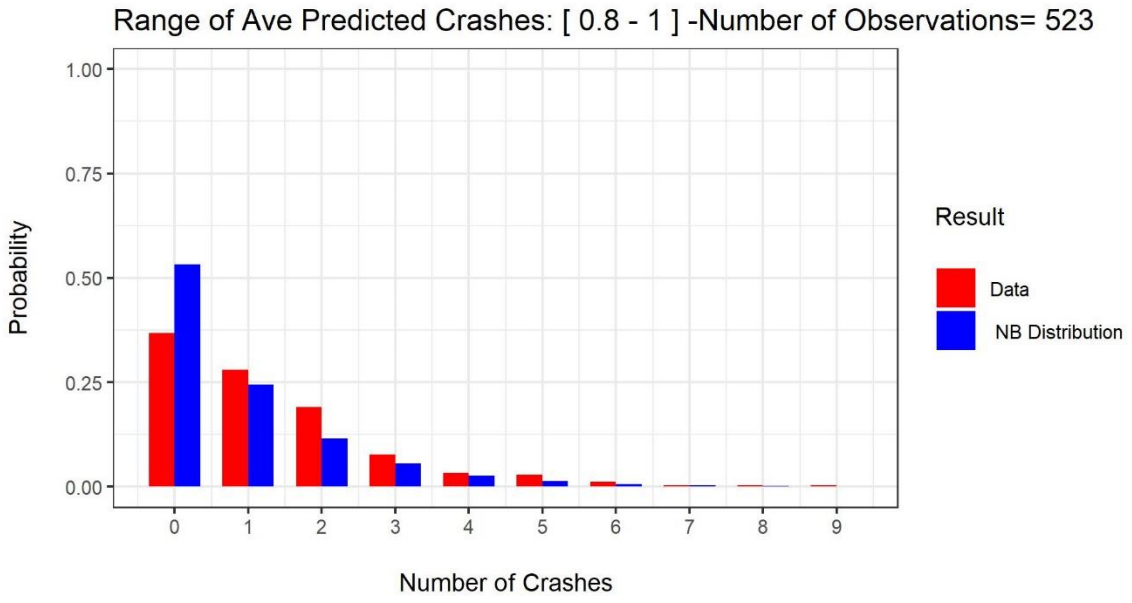


Figure 54 Comparison graph for the fifth range of average predicted crashes

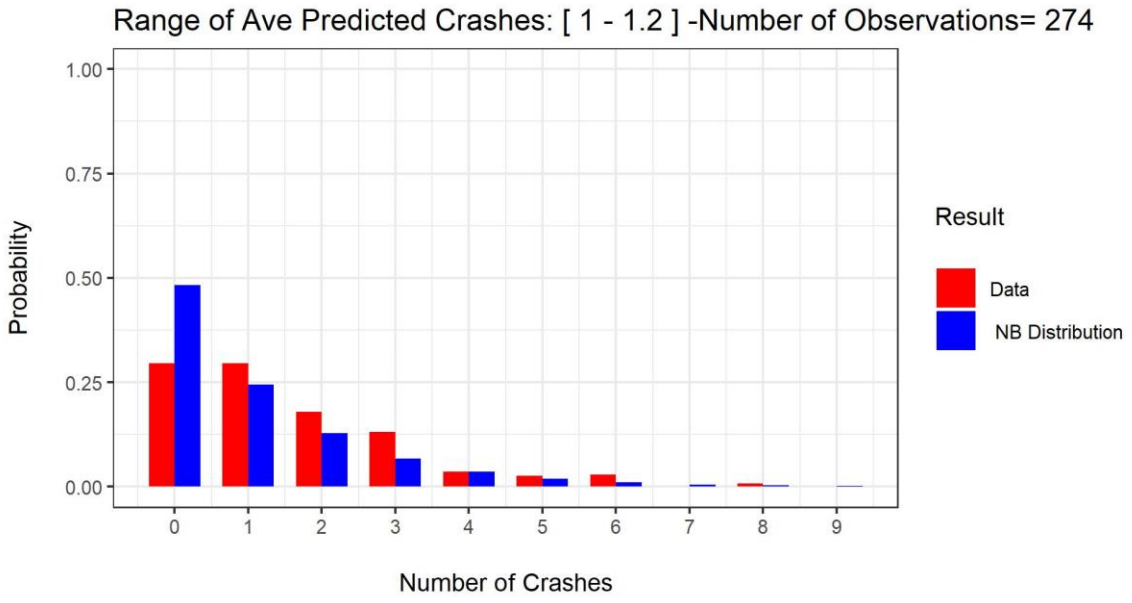


Figure 55 Comparison graph for the sixth range of average predicted crashes

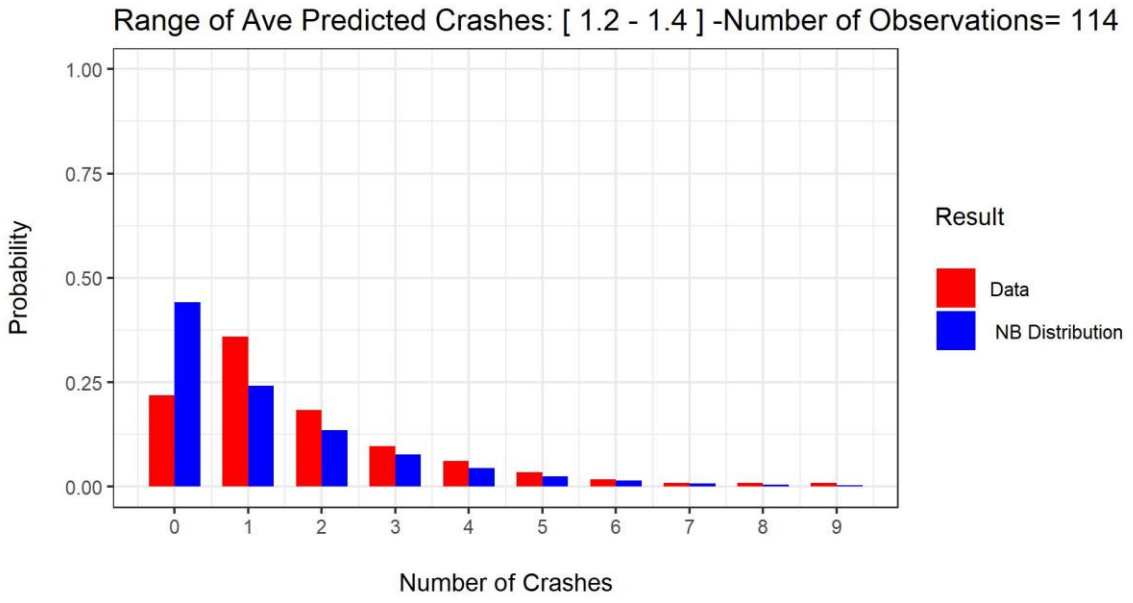


Figure 56 Comparison graph for the seventh range of average predicted crashes

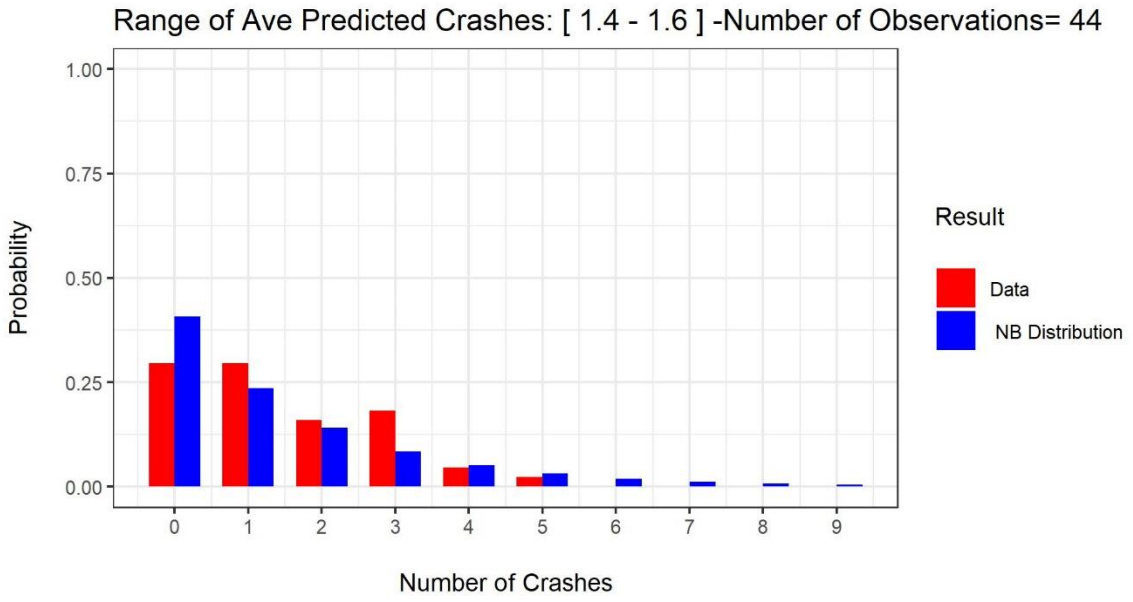


Figure 57 Comparison graph for the eighth range of average predicted crashes

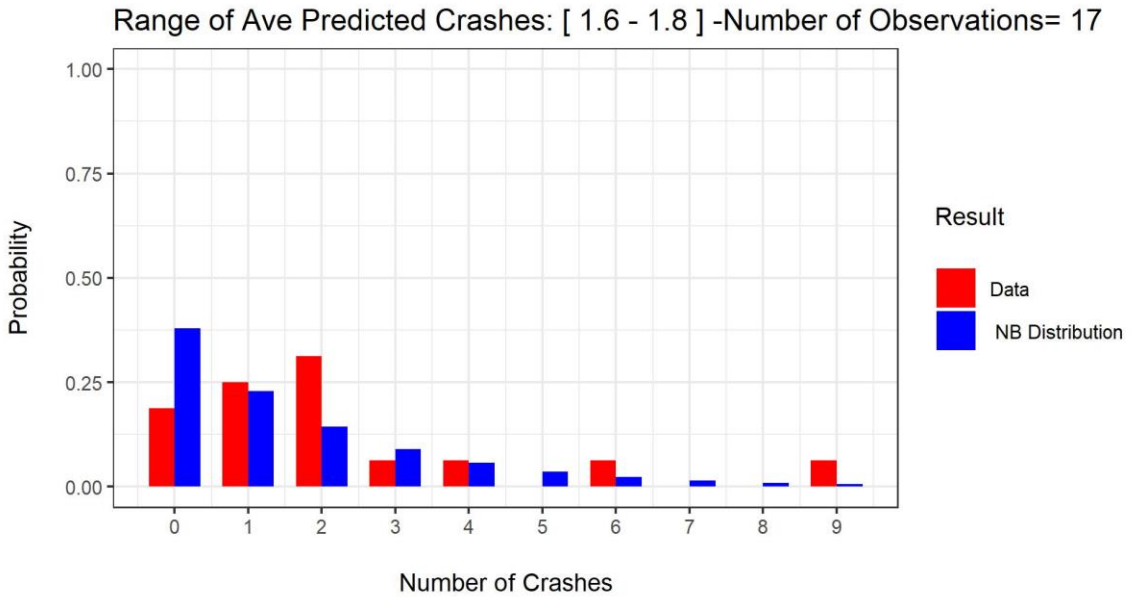


Figure 58 Comparison graph for the ninth range of average predicted crashes

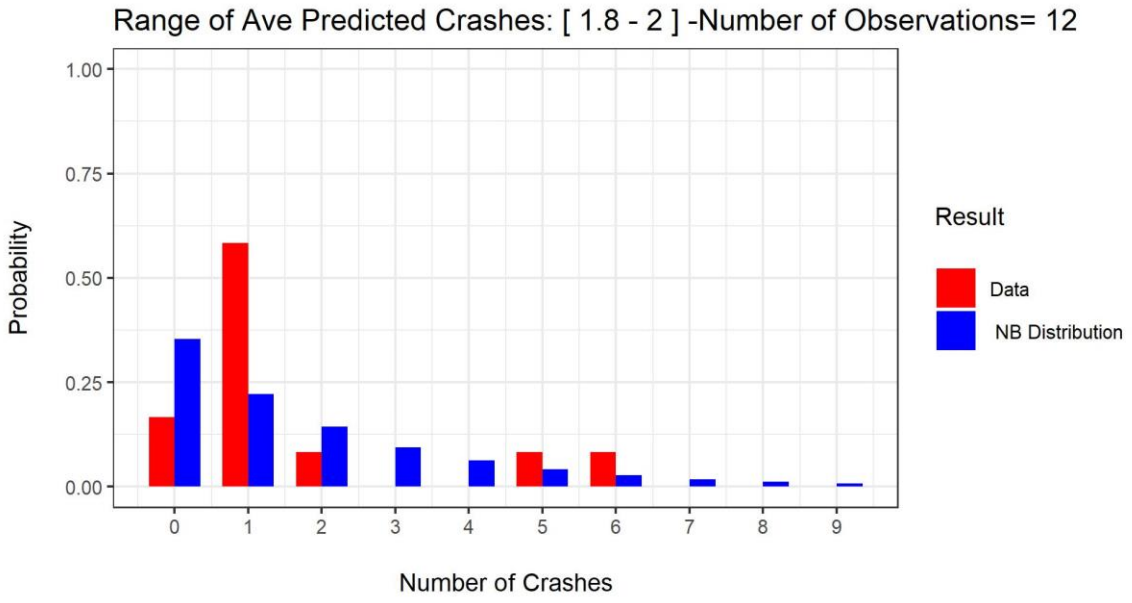


Figure 59 Comparison graph for the tenth range of average predicted crashes

Figure 60 presents a weighted average of the distributions presented for each range, where the number of observations provides the weights. As can be seen, the distributions are similar.

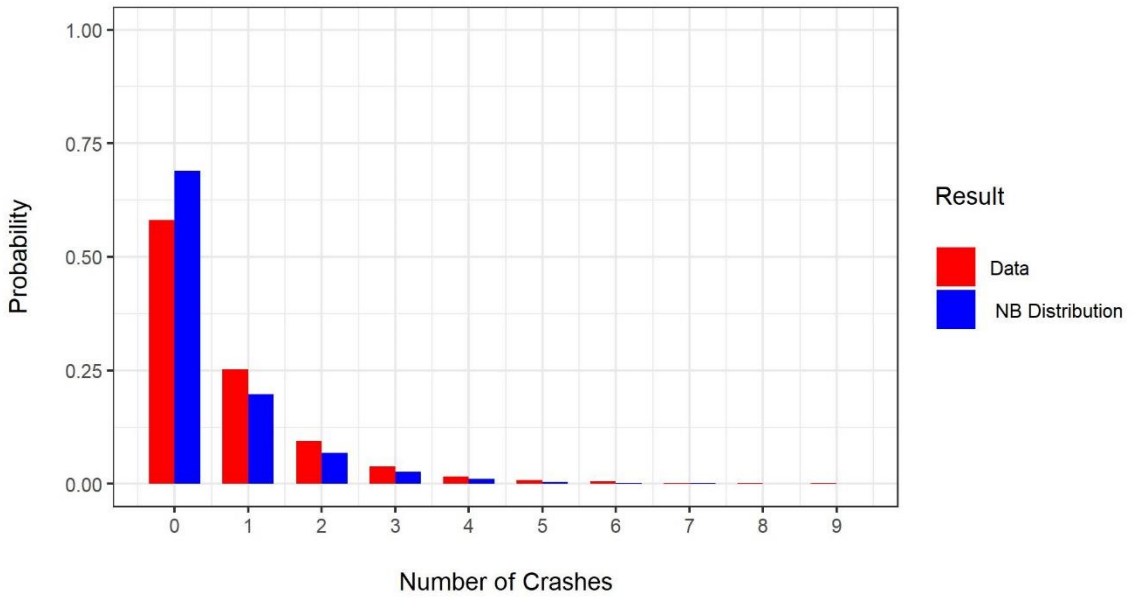


Figure 60 A weighted average of the distributions for all the ranges

Figure 61 presents the average predicted and observed crash rates for each of the Figures 50-59. Also, it shows the $R^2 = 0.67$ which is considerably less than previous models.

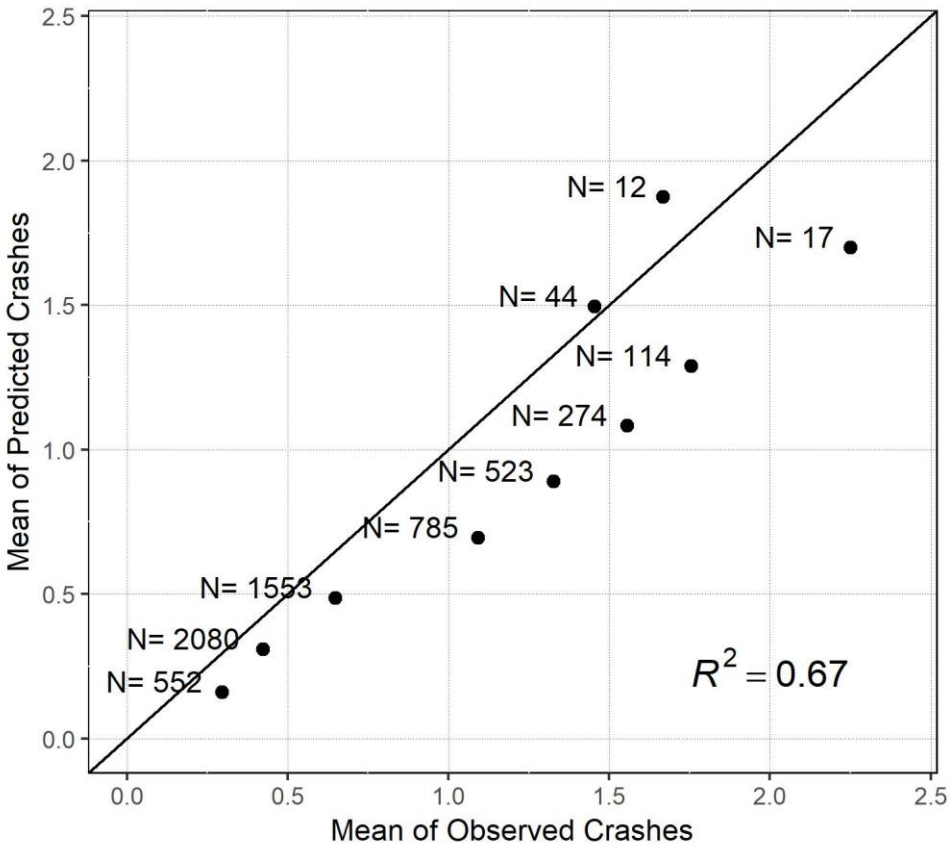


Figure 61 Observed versus predicted graph

4.3 Comparison of models

In this chapter, three statistical approaches were used to model the frequency of RwD crashes:

1. Negative binomial regression
2. Zero-inflated negative binomial regression
3. Mixed-effects negative binomial regression[1].

[1] The models are ordered with respect to their complexities or their statistical assumptions.

Each approach is based on different assumptions about the data generation process. The primary goal with using the different approaches was to explore their suitability for modeling the frequency of RwD crashes in Hawaii, and if possible to examine the potential superiority of one approach over another in terms of statistical results and other practical issues.

The estimation results with the three approaches were intuitively appealing in terms of the identification of statistically significant parameter estimates and their signs. In terms of observed versus predicted graphs (generated as described in section 4.2.1.2), the zero-inflated negative binomial model had the best fit. This is not surprising since the zero-inflated negative binomial model essentially combines two models (one to model the probability of zero inflation and another to model a frequency following a negative binomial probability distribution), which provides substantial flexibility to fit the data with a larger number of parameters. The negative binomial regression model (the most straightforward methodology here) had an acceptable fit as well. The statistical superiority of the zero-inflated negative binomial model over the negative binomial model was confirmed in Section 4.2.2.2 by a statistical test (i.e., Vuong test). As mentioned earlier, a considerable percentage of segments were observed to have zero-crashes during the study period. Therefore, the superiority of the zero-inflation model, which can handle the excessive number of zeros over the negative binomial model, was predictable. It is worth mentioning the two factors that contributed to the generation of excessive zeros in the database: 1-filtering out the crash database to the RwD crashes, and 2- selecting a short segment length to ensure the homogeneity of segments. However, the simplicity in the

interpretation of results to get practical inferences is another feature to consider in model selection. Although the zero-inflated model provides better statistics, its interpretation is more challenging than the negative binomial regression.

In addition to log-likelihood, the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) were utilized to evaluate the results of fitted models. The AIC and BIC were computed using equation 31 and 32, respectively.

$$AIC = -2LL + 2P \tag{31}$$

$$BIC = -2LL + P * \ln(n) \tag{32}$$

Where LL is the log-likelihood of the fitted model, P is the number of parameters used in the model, and n is the number of observations. As can be seen in the above equations, the AIC and BIC deal with the trade-off between the goodness of fit of the model and the simplicity of the model by introducing a penalty term for the number of parameters in the model. In general, a model with lower AIC and BIC values is preferred.

The mixed-effects negative binomial model (the most complex methodology here) had the best statistics of the three models (log-likelihood, AIC, and BIC). Table 7 compares these statistics for the three models. These were the result of the flexibility of the model to predict a separate set of coefficients for each observation. While the application of mixed-effects methodology was appealing (because it could consider the unobserved heterogeneity across the observations), it presents some challenges for making predictions since in practice only the fixed-effects can be used for this purpose because the contributions of the random-effects are typically unknown (unless the exact same dataset

with individual estimates of random effects is used). This issue could explain why the weakest results for observed versus predicted graphs belonged to the mixed-effects model.

Table 7 Model statistics comparison

	log-likelihood	**BIC**	**AIC**
Negative binomial model	-6835.3	13827.2	13707
Zero-inflated negative binomial model	-6759	13735.3	13568.3
Mixed-effects negative binomial model	-6722.7	13610.6	13483.4

Consideration of assumptions behind each methodology is crucial. An essential assumption of the zero-inflated negative binomial model is that a portion of zero-observed crashes is due to the inherent safety of the segment. However, the use of this methodology without having justifiable reasons that a segment can be purely safe (i.e., it is impossible to observe a crash on it) is questionable. While the logic behind this methodology is perfectly defensible in other fields of science[1], it does not seem to be free of challenges in crash frequency modeling. To a certain extent, the use of this methodology for modeling RwD crashes might be considered an artifact to fit a model to data that includes a

[1] This example is provided by IDRE-UCLA*(44)* "The state wildlife biologists want to model how many fish are being caught by fishermen at a state park. Some visitors do not fish, but there is no data on whether a person fished or not. Some visitors who did fish did not catch any fish so there are excess zeros in the data because of the people that did not fish". Note that there is no equivalence to persons that do not fish in RwD crash frequency modeling.

considerable portion of zero observations but without a strong relationship to the data generation process.

A common challenge with the mixed-effects methodology is the uncertainty that exists in the interpretation of results. Basically, the result of random-effects for each random parameter is a normal distribution with a mean of zero and an estimated standard deviation. In numerous studies*(16)(26)*, the estimated standard deviation of a random effect is close to or even higher than the corresponding estimated fixed-effects coefficient. Therefore, the interpretation of results for such parameters remains inconclusive since they would suggest that a unit increase in a variable would increase the number of crashes for a certain proportion of the segments and it would reduce the number of crashes for the complementary proportion of segments.

In terms of model parsimony[1], the negative binomial regression model is probably a good choice to be considered by decision-makers due to its simplicity.

In summary, the three estimated models provide similar results in terms of the variables identified as affecting the frequency of RwD crashes. From a statistical point of view, the mixed-effects model is the model with the best statistics (best log-likelihood, AIC, and BIC). On the other hand, in terms of fit (as described earlier), the zero-inflated negative binomial is better. However, there are concerns about the rationalization of the zero-inflation portion of the model. Finally, from the parsimony of view, the simpler

[1] A parsimonious model is a model that carry out a desired level of prediction with as few independent variables as possible.

negative binomial model is preferable. Constructive recommendations can be made based on the results of the three models to improve the safety of roads.

CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

This dissertation analyzed the frequency of RwD crashes, one of the deadliest type of crashes in the State of Hawaii, on TLTW state roads to investigate factors that affect their frequency. The analyses were based on ten years of RwD crashes data (including all severity levels) and roadway characteristics (e.g., traffic, geometry, and inventory databases) that can be aggregated at the segment level. Different methodologies were used to quantify the relationship between the frequency of RwD crashes and roadway characteristics. The most important conclusions derived from this research are as follows:

1. The segment length affects the model's estimations (i.e., the total number of statistically significant parameters and their estimated values). More specifically, the model estimation may not identify a parameter estimate statistically significant if the associated variable is highly dependent on the length of the segment. For example, the role of the average degree of curvature tends to wash out by extensive averaging for longer segment lengths.
2. The consideration of directional analysis improved the quality of the models in two ways: first, it allowed the assignment of head-on crashes based to the direction of the vehicles causing the crashes according to the police reports, which is more realistic for capturing the effects of geometric variables that may have actually contributed the most to the crash. With this directional assignment, it was possible to identify the contributing factors based on the direction of vehicles causing the

crashes[1]. An extreme example is that for the grade of the road that has a different value in each direction (a positive value for uphill and a negative value for downhills). The estimation results indicate that positive and negative grades have considerable different effects on the frequency of RwD crashes. Second, the directional analysis was the basis to include the general geometric environment of the analysis segment into the models.

3. The results indicate that the general geometric environment of the roadway portion where the segment was located affected the frequency of RwD crashes. For instance, the average degree of curvature before each segment affects the frequency of RwD crashes on the following segment. This means that for two similar segments, the frequency of RwD crashes are not equal if one is located on a winding road and the other segment is located right after a tangent road. This finding confirmed the importance of design consistency in highway design practices.

The following details were also considered to improve the quality of the models:

1. Assuming a quadratic form for the curvature and the grade. For instance, a quadratic form for a grade not only provided better statistics but also it provided a better clarification on the rate of changes in the frequency of RwD crashes in a possible range of the grade. In fact, the graph that depicted the contribution of the grade to the link function identified that RwD crashes increased at a higher rate on steeper negative grades (downhills) in comparison to the positive values (uphills).

[1] Because in most of the cases the roadway's characteristics are not the same in both directions.

2. Introducing an independent variable called “mean friction demand” based on the geometric relationship between the superelevation, speed and the degree of curvature.
3. Introducing the AADT both as a measure of exposure and separately as an independent variable to the RwD crashes.
4. Including the weighted standard deviation of some features like curvature not only improved the statistics of the model but also it compensated for some of the details that might be overlooked by just calculating the weighted average. For instance, the weighted standard deviation of curvature better represented the possible changes in the direction of consecutive curves in a segment.

In terms of model evaluation and the model selection, the following results were derived.

1. A new approach was introduced to evaluate the goodness of fit of generalized regression models based on their probabilistic nature. This approach is more consistent with theoretical concepts behind the negative binomial regression models, and it considers the probability density function of the count distribution. This approach was used to provide a better description of the goodness of fit to the data with respect to the approaches used in the literature. This is another tool to complement the typical evaluation of generalized regression models based on comparisons of the statistics of the models (e.g., their log-likelihood). The main contribution of this approach is that it permits a visual description of the model fit. However, equivalently to the use solely of R^2 for selecting linear models, its use for model selection is limited.

2. In terms of model evaluation, the mixed-effects model was found to provide the best statistics. However, its application on prediction was limited because only the fixed-effects could be used for the prediction. The zero-inflated model presented better statistics in comparison with the negative binomial regression; however, the assumptions behind this approach (i.e., the assumption that a segment can be inherently safe), and other complexities (e.g., the interpretations of results) lessen its practical application.

Lastly, the following results were derived in terms of the identification of factors that affect the frequency of RwD crashes[1].

1. The results identified a direct relationship between the rate[2] of RwD crashes and the mean friction demand, trucks percentage, absolute value of curvature, the standard deviation of curvature (environmental variable), IRI (environmental variable).
2. On the other hand, an increase in the following variables decreased the rate of RwD crashes: logarithm of AADT, shoulder width, grade, the proportion of the total length of sections with painted medians, lane width, the proportion of length of shoulder with asphalt concrete, curvature (environmental variable), the proportion of the total length of sections with guardrails, and the painted median width.

[1] Based on the negative binomial regression model by considering the segment length equals to 0.2-mile

[2] Rate models were explained in Section 2.2.

5.2 Practical implementations

The results provided some insight to select countermeasures to reduce RwD crashes either by keeping the vehicle on the roadway or by reducing the potential for crashes when vehicles do leave the road. The results suggest that the following countermeasures may reduce the frequency of RwD crashes: Using a high friction surface treatment on segments with high mean friction demand to provide a higher safety margin between friction demand and friction supply; to increase the superelevation on curves with high friction demand to decrease it; widening lanes, shoulders, and the painted median; paving the shoulders with asphalt concrete, installing guardrails, improving the roadway geometrics (e.g., curve delineation, grades); resurfacing the pavement to improve the IRI[1].

Based on the predictions with the models developed in this study (i.e., the mean value for the frequency of RwD crashes on each segment), the segments with a higher likelihood of the RwD crashes may be prioritized for further safety projects due to the limited resources. Decision-makers may consider the results to reduce the total number of RwD crashes effectively on TLTW state roads by prioritizing the locations and the types of countermeasures.

5.3 Model limitations

Even though it is believed that the crash frequency models developed in this study represent an improvement from other models that were developed to study the frequency

[1] It also improves the ride quality.

of RwD crashes, it also has some limitations that are important to point out:

1. A common limitation of crash frequency models is that they are developed based on police crash reports. Hence, it is highly probable that the total number of RwD crashes is more than the reported crashes. Likewise, it is most likely that the total number of property damage only (PDO) crashes and non-injury crashes are underestimated. Since the models developed in this dissertation were based on the police crash reports, they share this limitation.
2. The unit of analysis is the roadway segment in crash frequency modeling. Therefore, only information relating to the roadway segments can be used to predict the crashes on segments. Therefore, the findings of this methodology emphasize the roadway characteristics incorporated in highway design (e.g., traffic, geometry, and inventory databases). The findings of this dissertation are mostly beneficial in promoting a safer highway design to reduce the frequency of RwD crashes. It is worth to mention that the attributes of drivers, passengers, vehicles, and the weather undoubtedly can affect the frequency of RwD crashes (even more so that the roadway features), but this information cannot be incorporated into this methodology other than by segmenting the data into data sets with different conditions and estimating separate models. The attributes of drivers, passengers, vehicles, and the weather information are more amenable for a separate type of analysis (i.e., crash severity models) that predict the severity of crashes. However, such an analysis is beyond the scope of this dissertation.

5.4 Future research

The models presented in this dissertation captured the effects of roadway characteristics on the frequency of RwD crashes; however, additional benefits can be realized through the following recommendations:

1. Data improvements:

   Improving the quality of current models by adding some important predictor variables unavailable in this study is recommended. For instance, adding surface friction as an independent variable may improve the results. For this dissertation, such information was not available, so an attempt was made to calculate the mean friction demand instead based on the other geometric features of the roads. Using the measured average surface friction (friction supply) for model estimation is expected to improve the model by capturing the difference between friction demand and friction supply.

   Another example is capturing the effects of existing low-cost safety improvements such as rumble strips and chevron signs. Currently, such safety improvements are present in many of the segments in the database, but they were not considered because of their location or year of construction were not known. The information for some segments in the existing dataset (possibly expanded with additional years) could be split into before and after application of a safety treatment. Re-estimation of such a model could provide valuable information to judge the effectiveness of such treatments. This is the main reason that the models were estimated with an offset for the number of years even though for all the sections in this study exactly ten years of data were available. If the dataset is modified as indicated above, the number of years for some observations

would be different, but this can be easily accounted for with the use of the offset for the number of years.

2. Crash severity analysis:

   Developing crash severity models based on the same data to complement the outcomes of the current study. The severity models reveal the determinant factors that affect the severity of RwD crashes. The primary intention is to identify the factors that reduce the total number of RwD crashes with incapacitating injuries or deaths.

3. Driving Under the Influence (DUI)

   It is widely accepted that driving under the influence (e.g., alcohol and/or drugs) is one of the primary reasons of crashes. While improving the highways' safety to reduce the total number of crashes for all the users (i.e., including impaired drivers) might be of interest, it might not be a cost-effective strategy because of the limitations of the safety budget. Alternatively, roads can be designed and improved with required standards for the majority of users (excluding the impaired drivers) and meanwhile, part of the safety budget could be allocated to educate drivers about the consequences of driving under the influence. Therefore, it is recommended to split the dataset into two DUI related and non-DUI related RwD crashes to highlight the effects of DUI. Consequently, two separate crash frequency models can be developed, and statistical inferences based on separately developed models can be derived to help decision-makers for further safety improvements.

4. Segment length analysis:

   Separate crash frequency models were developed to emphasize the importance of segment length in crash frequency modeling. An interesting topic to continue this endeavor is to find a robust methodology to identify the best segment length.

5. The general geometric environment of the roadway:

   The importance of general geometric environment was investigated in this dissertation, and it was found that the information for 1.2-mile (~ 2.0-km) before each segment affected the frequency of RwD crashes. However, through a well-established methodology, there might be an optimum length before each segment that fully reflects the role of the general geometric environment.

REFERENCES

1. Federal Highway Administration. Roadway Departure (RwD) Strategic Plan. https://safety.fhwa.dot.gov/roadway_dept/docs/rwd_strategic_plan_version2013.pdf. Accessed Mar. 1, 2018.

2. Neuman, T. R., R. Pfefer, K. L. Slack, K. K. Hardy, F. Council, H. McGee, L. Prothe, and K. Eccles. *A guide for addressing run-off-road collisions*. 2003.

3. Neuman, T., R. Pfefer, K. L. Slack, H. McGee, L. Prothe, K. Eccles, and F. Council. *NCHRP REPORT 500: Guidance for Implementation of the AASHTO Strategic Highway Safety Plan.Volume 4: A Guide for Addressing Head-On Collisions*. 2003.

4. AASHTO. AASHTO Strategic Highway Safety Plan (SHSP). *Washington, DC*, 2005, pp. 24–31.

5. Federal Highway Administration. Roadway Departure Safety. https://safety.fhwa.dot.gov/roadway_dept/. Accessed Mar. 1, 2018.

6. Bonneson, J. A. Highway safety manual. *Washington, DC: American Association of State Highway and Transportation Officials*, 2010.

7. Preston, H., R. Storm, J. D. Bennett, and E. Wemple. Systemic Safety Project Selection Tool, Federal Highway Administration, US Department of Transportation. 2013, p. 100.

8. Agresti, A. *Foundations of linear and generalized linear models*. John Wiley & Sons, 2015.

9. Mannering, F. L., and C. R. Bhat. Analytic methods in accident research: Methodological frontier and future directions. *Analytic Methods in Accident Research*, Vol. 1, Jan. 2014, pp. 1–22.

10. Miaou, S.-P. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. *Accident Analysis & Prevention*, Vol. 26, No. 4, 1994, pp. 471–482.

11. Shankar, V., F. Mannering, and W. Barfield. Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. *Accident Analysis & Prevention*, Vol. 27, No. 3, Jun. 1995, pp. 371–389.

12. Lee, J., and F. Mannering. Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis. *Accident Analysis & Prevention*, Vol. 34, No. 2, Mar. 2002, pp. 149–161.

13. Lord, D., S. D. Guikema, and S. R. Geedipally. Application of the Conway-Maxwell-Poisson generalized linear model for analyzing motor vehicle crashes. *Accident Analysis and Prevention*, Vol. 40, No. 3, May 2008, pp. 1123–1134.

14. Hauer, E. *The art of regression modeling in road safety*. Springer, 2016.

15. Geedipally, S. R., D. Lord, and S. S. Dhavala. The negative binomial-Lindley generalized linear model: Characteristics and application using crash data. *Accident Analysis and Prevention*, Vol. 45, Mar. 2012, pp. 258–265.

16. Anastasopoulos, P. C., and F. L. Mannering. A note on modeling vehicle accident frequencies with random-parameters count models. *Accident; analysis and*

prevention, Vol. 41, No. 1, Jan. 2009, pp. 153–9.

17. Faraway, J. J. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. CRC press, 2016.

18. Mannering, F. L., V. Shankar, and C. R. Bhat. Unobserved heterogeneity and the statistical analysis of highway accident data. *Analytic Methods in Accident Research*, Vol. 11, Sep. 2016, pp. 1–16.

19. Lord, D., and F. Mannering. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*, Vol. 44, No. 5, Jun. 2010, pp. 291–305.

20. Washington, S. P., M. G. Karlaftis, and F. L. Mannering. *Statistical and econometric methods for transportation data analysis*. CRC press, 2010.

21. Winkelmann, R. *Econometric analysis of count data*. Springer Science & Business Media, 2008.

22. Lord, D., S. P. Washington, and J. N. Ivan. Poisson, poisson-gamma and zero-inflated regression models of motor vehicle crashes: Balancing statistical fit and theory. *Accident Analysis and Prevention*, Vol. 37, No. 1, Jan. 2005, pp. 35–46.

23. Lord, D., S. Washington, and J. N. Ivan. Further notes on the application of zero-inflated models in highway safety. *Accident Analysis and Prevention*, Vol. 39, No. 1, Jan. 2007, pp. 53–57.

24. Greene, W. H. *Econometric analysis*. Pearson Education India, 2003.

25. Anastasopoulos, P. C., F. L. Mannering, V. N. Shankar, and J. E. Haddock. A study of factors affecting highway accident rates using the random-parameters tobit model. *Accident; analysis and prevention*, Vol. 45, Mar. 2012, pp. 628–33.

26. Rusli, R., M. M. Haque, M. King, and W. S. Voon. Single-vehicle crashes along rural mountainous highways in Malaysia: An application of random parameters negative binomial model. *Accident Analysis & Prevention*, Vol. 102, 2017, pp. 153–164.

27. Venkataraman, N., G. F. Ulfarsson, and V. N. Shankar. Random parameter models of interstate crash frequencies by severity, number of vehicles involved, collision and location type. *Accident Analysis & Prevention*, Vol. 59, 2013, pp. 309–318.

28. Pinheiro, J. C., and D. M. Bates. Mixed-Effects Models in S and S-plus. *Statistics and computing*, 1978.

29. Zuur, A. F., E. N. Ieno, N. J. Walker, A. A. Saveliev, and G. M. Smith. Mixed effects models and extensions in ecology with R. Gail M, Krickeberg K, Samet JM, Tsiatis A, Wong W, editors. *New York, NY: Spring Science and Business Media*, 2009.

30. R Core Team. R: A Language and Environment for Statistical Computing. http://www.r-project.org/.

31. Peng, Y., S. R. Geedipally, and D. Lord. Investigating the Effect of Roadside features on Single-Vehicle Roadway Departure Crashes on Rural Two-Lane Roads. 2012.

32. Kim, K., P. Pant, and E. Yamashita. Accidents and accessibility: Measuring influences of demographic and land use variables in Honolulu, Hawaii. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2147, 2010, pp. 9–17.

33. Hashemi, M., and A. R. Archilla. Potential Factors Affecting Roadway Departure Crashes in Oahu, Hawaii. *ITE Western District Annual Meeting, Albuquerque*, 2016, p. 10.

34. Hashemi, M., and A. R. Archilla. Exploratory Analysis of Roadway Departure Crashes Contributing Factors Based on Classification and Regression Trees. *ITE Western District Annual Meeting, San Diego*, 2017.

35. Honolulu Land Information Systems. http://gis.hicentral.com/data.html.

36. Milton, J., and F. Mannering. The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies. *Transportation*, Vol. 25, No. 4, 1998, pp. 395–413.

37. Mallick, R. B., and T. El-Korchi. *Pavement engineering: principles and practice*. CRC Press, 2017.

38. Ripley, B. MASS: Support Functions and Datasets for Venables and Ripley's MASS. https://cran.r-project.org/package=MASS.

39. Zhang, C., and J. Ivan. Effects of geometric characteristics on head-on crash incidence on two-lane roads in Connecticut. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1908, 2005, pp. 159–164.

40. Fearon, S. J. and A. T. and A. Z. and C. M. and J. pscl: Political Science Computational Laboratory. https://cran.r-project.org/package=pscl.

41. Walker, D. B. and M. M. and B. B. and S. lme4: Linear Mixed-Effects Models using "Eigen" and S4. https://cran.r-project.org/package=lme4.

42. Wolfinger, R. Laplace's approximation for nonlinear mixed models. *Biometrika*, Vol. 80, No. 4, 1993, pp. 791–795.

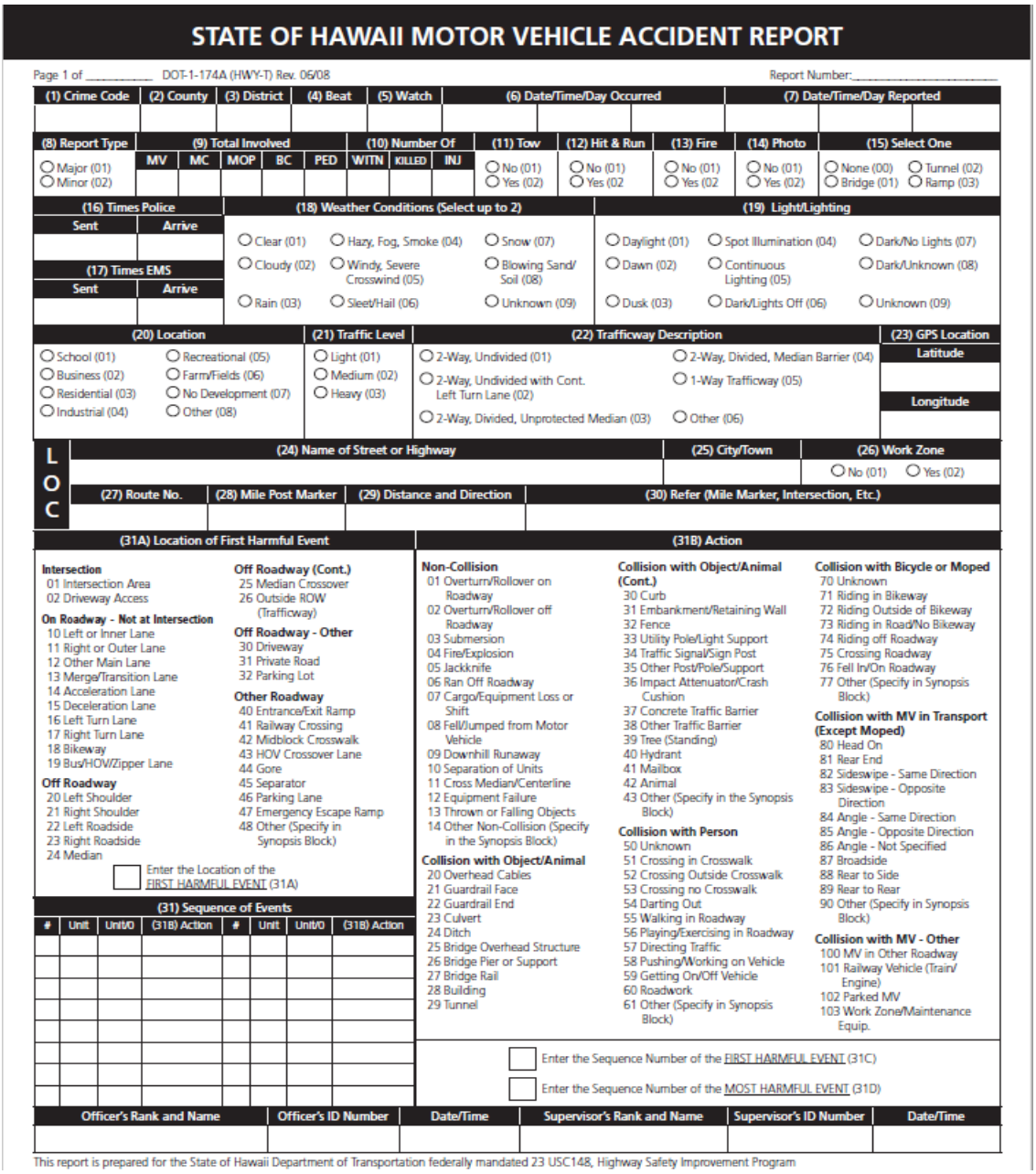
43. Federal Highway Administration. KABCO Injury Classification Scale and Definitions. https://safety.fhwa.dot.gov/hsip/spm/conversion_tbl/pdfs/kabco_ctable_by_state.pdf. Accessed Apr. 1, 2019.

44. UCLA: IDRE (Institute for Digital Research Education). ZERO-INFLATED NEGATIVE BINOMIAL REGRESSION. https://stats.idre.ucla.edu/r/dae/zinb/. Accessed Apr. 1, 2019.

APPENDIX

A. State of Hawaii motor vehicle accident report form

The state of Hawaii motor vehicle accident report form is provided in this section.



STATE OF HAWAII MOTOR VEHICLE ACCIDENT REPORT

Page 1 of _____ DOT-1-174A (HWY-T) Rev. 06/08 Report Number:_____

(1) Crime Code	(2) County	(3) District	(4) Beat	(5) Watch	(6) Date/Time/Day Occurred	(7) Date/Time/Day Reported

(8) Report Type	(9) Total Involved					(10) Number Of			(11) Tow	(12) Hit & Run	(13) Fire	(14) Photo	(15) Select One
	MV	MC	MOP	BC	PED	WITN	KILLED	INJ					
○ Major (01) ○ Minor (02)									○ No (01) ○ Yes (02)	○ No (01) ○ Yes (02	○ No (01) ○ Yes (02	○ No (01) ○ Yes (02)	○ None (00) ○ Bridge (01) ○ Tunnel (02) ○ Ramp (03)

(16) Times Police		(17) Times EMS		(18) Weather Conditions (Select up to 2)			(19) Light/Lighting		
Sent	Arrive	Sent	Arrive						
				○ Clear (01)	○ Hazy, Fog, Smoke (04)	○ Snow (07)	○ Daylight (01)	○ Spot Illumination (04)	○ Dark/No Lights (07)
				○ Cloudy (02)	○ Windy, Severe Crosswind (05)	○ Blowing Sand/ Soil (08)	○ Dawn (02)	○ Continuous Lighting (05)	○ Dark/Unknown (08)
				○ Rain (03)	○ Sleet/Hail (06)	○ Unknown (09)	○ Dusk (03)	○ Dark/Lights Off (06)	○ Unknown (09)

(20) Location		(21) Traffic Level	(22) Trafficway Description		(23) GPS Location
					Latitude
○ School (01)	○ Recreational (05)	○ Light (01)	○ 2-Way, Undivided (01)	○ 2-Way, Divided, Median Barrier (04)	
○ Business (02)	○ Farm/Fields (06)	○ Medium (02)	○ 2-Way, Undivided with Cont. Left Turn Lane (02)	○ 1-Way Trafficway (05)	Longitude
○ Residential (03)	○ No Development (07)	○ Heavy (03)			
○ Industrial (04)	○ Other (08)		○ 2-Way, Divided, Unprotected Median (03)	○ Other (06)	

LOC	(24) Name of Street or Highway	(25) City/Town	(26) Work Zone
			○ No (01) ○ Yes (02)

(27) Route No.	(28) Mile Post Marker	(29) Distance and Direction	(30) Refer (Mile Marker, Intersection, Etc.)

(31A) Location of First Harmful Event

Intersection
01 Intersection Area
02 Driveway Access

On Roadway - Not at Intersection
10 Left or Inner Lane
11 Right or Outer Lane
12 Other Main Lane
13 Merge/Transition Lane
14 Acceleration Lane
15 Deceleration Lane
16 Left Turn Lane
17 Right Turn Lane
18 Bikeway
19 Bus/HOV/Zipper Lane

Off Roadway
20 Left Shoulder
21 Right Shoulder
22 Left Roadside
23 Right Roadside
24 Median

Off Roadway (Cont.)
25 Median Crossover
26 Outside ROW (Trafficway)

Off Roadway - Other
30 Driveway
31 Private Road
32 Parking Lot

Other Roadway
40 Entrance/Exit Ramp
41 Railway Crossing
42 Midblock Crosswalk
43 HOV Crossover Lane
44 Gore
45 Separator
46 Parking Lane
47 Emergency Escape Ramp
48 Other (Specify in Synopsis Block)

☐ Enter the Location of the FIRST HARMFUL EVENT (31A)

(31) Sequence of Events

#	Unit	Unit/O	(31B) Action	#	Unit	Unit/O	(31B) Action

(31B) Action

Non-Collision
01 Overturn/Rollover on Roadway
02 Overturn/Rollover off Roadway
03 Submersion
04 Fire/Explosion
05 Jackknife
06 Ran Off Roadway
07 Cargo/Equipment Loss or Shift
08 Fell/Jumped from Motor Vehicle
09 Downhill Runaway
10 Separation of Units
11 Cross Median/Centerline
12 Equipment Failure
13 Thrown or Falling Objects
14 Other Non-Collision (Specify in the Synopsis Block)

Collision with Object/Animal
20 Overhead Cables
21 Guardrail Face
22 Guardrail End
23 Culvert
24 Ditch
25 Bridge Overhead Structure
26 Bridge Pier or Support
27 Bridge Rail
28 Building
29 Tunnel

Collision with Object/Animal (Cont.)
30 Curb
31 Embankment/Retaining Wall
32 Fence
33 Utility Pole/Light Support
34 Traffic Signal/Sign Post
35 Other Post/Pole/Support
36 Impact Attenuator/Crash Cushion
37 Concrete Traffic Barrier
38 Other Traffic Barrier
39 Tree (Standing)
40 Hydrant
41 Mailbox
42 Animal
43 Other (Specify in the Synopsis Block)

Collision with Person
50 Unknown
51 Crossing in Crosswalk
52 Crossing Outside Crosswalk
53 Crossing no Crosswalk
54 Darting Out
55 Walking in Roadway
56 Playing/Exercising in Roadway
57 Directing Traffic
58 Pushing/Working on Vehicle
59 Getting On/Off Vehicle
60 Roadwork
61 Other (Specify in Synopsis Block)

Collision with Bicycle or Moped
70 Unknown
71 Riding in Bikeway
72 Riding Outside of Bikeway
73 Riding in Road/No Bikeway
74 Riding off Roadway
75 Crossing Roadway
76 Fell In/On Roadway
77 Other (Specify in Synopsis Block)

Collision with MV in Transport (Except Moped)
80 Head On
81 Rear End
82 Sideswipe - Same Direction
83 Sideswipe - Opposite Direction
84 Angle - Same Direction
85 Angle - Opposite Direction
86 Angle - Not Specified
87 Broadside
88 Rear to Side
89 Rear to Rear
90 Other (Specify in Synopsis Block)

Collision with MV - Other
100 MV in Other Roadway
101 Railway Vehicle (Train/ Engine)
102 Parked MV
103 Work Zone/Maintenance Equip.

☐ Enter the Sequence Number of the FIRST HARMFUL EVENT (31C)
☐ Enter the Sequence Number of the MOST HARMFUL EVENT (31D)

Officer's Rank and Name	Officer's ID Number	Date/Time	Supervisor's Rank and Name	Supervisor's ID Number	Date/Time

This report is prepared for the State of Hawaii Department of Transportation federally mandated 23 USC148, Highway Safety Improvement Program

Page _____ of _____
DOT-1-1748 (HWY-T) Rev. 06/08

STATE OF HAWAII MOTOR VEHICLE ACCIDENT REPORT

Report Number:_____

UNIT INFORMATION

(32) Unit No.	(33) No. of Occ.

(34) Unit Class

- ○ Passenger Car (01)
- ○ Passenger Van (02)
- ○ Pickup Truck (03)
- ○ SUV/MPVH (04)
- ○ Cargo Van < 10,001 lbs. (05)
- ○ Other Truck < 10,001 lbs. (06)
- ○ Truck > 10,000 lbs. (07)
- ○ Transit Bus (08)
- ○ School Bus (09)
- ○ Other Bus (10)
- ○ Motorcycle (11)
- ○ Motor Scooter (12)
- ○ Moped (13)
- ○ Bicycle (14)
- ○ Pedestrian (15)
- ○ Maint./Construct. Equipment (16)
- ○ Farm Vehicle/Equipment (17)
- ○ Motor Coach (18)
- ○ Motor Home (19)
- ○ Recreational Vehicle (20)
- ○ Other (21)
- ○ Unknown (22)

(35) Race

- ○ White (01)
- ○ Black (02)
- ○ American Indian (03)
- ○ Chinese (04)
- ○ Japanese (05)
- ○ Korean (06)
- ○ Puerto Rican (07)
- ○ Hawaiian (08)
- ○ Samoan (09)
- ○ Tongan (10)
- ○ Vietnamese (11)
- ○ Filipino (12)
- ○ Unknown (13)
- ○ Other (14)

(36) Last Name	(37) First Name	(38) MI	(39) Sex	(40) DOB
			○ M (01) ○ F (02)	

(41) Street No.	(42) Street Name	(43) St., Pl., Blvd., Etc.	(44) Apt/Suite Number

(45) City	(46) State	(47) Zip Code	(48) Home Phone Number
			( ) -

(49) Occupation

- ○ Unemployed (00)
- ○ U.S. Army (01)
- ○ U.S. Navy (02)
- ○ U.S. Air Force (03)
- ○ U.S. Marines (04)
- ○ U.S. Coast Guard (05)
- ○ Other Military (06)
- ○ Fed. Govt. Civ. (07)
- ○ State Govt. (08)
- ○ County Govt. (09)
- ○ Foreign Govt./Civ. (10)
- ○ Retired (11)
- ○ Student - Elem. (12)
- ○ Student - Inter. (13)
- ○ Student - H.S. (14)
- ○ Student - Col. (15)
- ○ U.S. Tourist (16)
- ○ Foreign Tourist (17)
- ○ Police Officer (18)
- ○ Other (19)
- ○ Not Stated (20)

(50) Employer/Company Name

(51) Work Phone Number	(52) Other Phone/Pager Number
( ) -	( ) -

(53) Driver's License Number	(54) St./Juris.	(55) Class	(56) Restrict.	(57) Endorse.

(58) CDL Type

- ○ Non-CDL (01)
- ○ Non-CDL/Restricted (02)
- ○ CDL (03)

(59) Driver's License Status

- ○ Valid (01)
- ○ Not Licensed (02)
- ○ Canceled (03)
- ○ Denied (04)
- ○ Expired (05)
- ○ Revoked (06)
- ○ Suspended (07)
- ○ Provisional (08)
- ○ Permit (09)
- ○ Disqualified [CDL] (10)

(85) SFST GIVEN

- ○ No (01)
- ○ Yes (02)
- ○ Refused (03)

(86) SFST Results

- ○ Passed (01)
- ○ Failed (02)
- ○ Does Not Apply (03)

(87) Alcohol Test Results

(87A) Status	(87B) Type	(87C) Results
○ None (00)	○ Blood (01)	○ Value (01)
○ Refused (01)	○ Breath (02)	
○ Given (02)	○ Other (03)	○ Pending (02)

(88) Drug Test Results

(88A) Status	(88B) Type	(88C) Results
○ None (00)	○ Blood (01)	○ Positive (01)
○ Refused (01)	○ Urine (02)	○ Negative (02)
○ Given (02)	○ Other (03)	○ Pending (03)

(60) Insurance Policy Number	(61) Exp. Date	(62) Insurance Carrier

(63) Registered Owner Name	(64) Phone Number
	( ) -

(65) Str. No.	(66) Street Name	(67) St., Pl.	(68) Ste. #

(69) City	(70) State	(71) Zip Code

(72) Vehicle Body Type

- ○ 2-DSD (01)
- ○ 4-DSD (02)
- ○ 2-DCV (03)
- ○ 2-DSW (04)
- ○ 4-DSW (05)
- ○ P/U Truck (06)
- ○ SUV/MPVH (07)
- ○ Van (08)
- ○ Truck (09)
- ○ Bus (10)
- ○ PCMC (11)
- ○ M-Scooter (12)
- ○ Moped (13)
- ○ Bicycle (14)
- ○ Other (15)

(73) Vehicle Year	(74) Veh. Color (Top/Bottom)	(75) Vehicle Make	(76) Vehicle Model	(77) Lic. Plate No.	(78) Trailer Plate	(79) Lic. Plate St.

(80) Vehicle VIN Number	(81) Emer. Veh. In Use	(82) Vehicle Stolen
	○ No (01) ○ Yes (02)	○ No (01) ○ Yes (02)

(83) Special Use

- ○ None (00)
- ○ Driver Trng. (01)
- ○ Construction/ Maintenance (02)
- ○ Taxi (03)
- ○ Fire Truck (04)
- ○ Tow Truck (05)
- ○ Ambulance (06)
- ○ Police-On Duty (07)
- ○ Police-Off Duty (08)
- ○ Military (09)
- ○ Government (10)
- ○ Farm Use (11)
- ○ U-Drive (12)
- ○ School Bus (13)
- ○ Other Bus (14)
- ○ Other (15)

(84) Trailer/Cargo Type

- ○ None (00)
- ○ Boat (01)
- ○ Flatbed (02)
- ○ Horse (03)
- ○ Livestock (04)
- ○ House (05)
- ○ Van/Encl. Box (06)
- ○ Dump (07)
- ○ Veh. Tow Veh. (08)
- ○ Other (09)
- ○ N/A (10)

Officer's Initials:

Supervisor's Initials:

Page _____ of _____
DOT-1-174C (HWY-T) Rev. 06/08

STATE OF HAWAII MOTOR VEHICLE ACCIDENT REPORT

Report Number:_____

Unit No.

UNIT INFORMATION (Cont.)

(89) Citations

Citation Number	Offense Code (H.R.S./R.O. Section No.)

(90) Est. Damages
- ○ $3,000 or Greater (01)
- ○ Less than $3,000 (02)

(91) Extent of Damage
- ○ None (00)
- ○ Functional (02)
- ○ Minor (01)
- ○ Disabling (03)

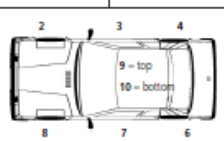
(92) Is this a CMV or Other QUALIFYING Vehicle?
- ○ No (01)
- ○ Yes (02)

If yes, go to CMV SUPPLEMENT

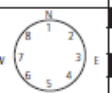
(95A) Object (1) Struck/Damage Description

(96A) Object (2) Struck/Damage Description

(93) Using the Diagram to the Right, Indicate Initial Impact Point in block below:



Circle Damaged Areas



(94) Direction

From	To

(95B) (Object 1) Owner's Name

(96B) (Object 2) Owner's Name

(95C) (Object 1) Owner's Phone Number

( ) -

(96C) (Object 2) Owner's Phone Number

( ) -

(95D) Estimated Damages to Object 1
- ○ $3,000 or Greater (01)
- ○ Less than $3,000 (02)

(96D) Estimated Damages to Object 2
- ○ $3,000 or Greater (01)
- ○ Less than $3,000 (02)

(97) Motor Vehicle Maneuver/Action
- ○ Straight Ahead (01)
- ○ Changing Lanes (02)
- ○ Merging (03)
- ○ Overtaking/ Passing (04)
- ○ Slowing/Stopping (05)
- ○ Backing (06)
- ○ Parking (07)
- ○ Parked (08)
- ○ Start from Parked (09)
- ○ Stopped in Traffic (10)
- ○ Start in Traffic (11)
- ○ Right Turn on Red (12)
- ○ Turning Right (13)
- ○ Turning Left (14)
- ○ U-Turn (15)
- ○ Entering Traffic (16)
- ○ Negotiating a Curve (17)
- ○ Other (18)

(98) Reason for Maneuver
- ○ Intended Maneuver (01)
- ○ Traffic Controls (02)
- ○ Mechanical Failure (03)
- ○ Avoid Other Vehicle (04)
- ○ Avoid Pedestrian (05)
- ○ Avoid Bicycle (06)
- ○ Avoid Obj./ Animal (07)
- ○ Avoid Prior MVA (08)
- ○ Other (09)

(99) Traffic Control Device Type
- ○ No Controls (00)
- ○ Traffic Signal (01)
- ○ Stop Sign (02)
- ○ Yield Sign (03)
- ○ Flashing Red (04)
- ○ Flashing Amber (05)
- ○ Person (06)
- ○ School Zone Sign/ Device (07)
- ○ Warning Sign (08)
- ○ Railway X-ing Device (09)
- ○ Other (10)

(100) Traffic Control Condition
- ○ Functioning Properly (01)
- ○ Knocked Down (02)
- ○ Obscured (03)
- ○ Red Malfunction (04)
- ○ Yellow Malfunction (05)
- ○ Green Malfunction (06)
- ○ Arrow Malfunction (07)
- ○ Lights Not Changing (08)
- ○ Other Malfunction (09)

(101) Guidance/Pavement Markings

	Lft	Rgt		Lft	Rgt
None (00)	○	○	No Passing, Yellow (06)	○	○
Solid Yellow (01)	○	○	Curb/Median, Etc. (07)	○	○
Skip-Dash Yellow (02)	○	○	Bikeway Marking (08)	○	○
Solid White (03)	○	○	Crosswalk Marking (09)	○	○
Skip-Dash White (04)	○	○	Turn Lane (10)	○	○
Solid Double Yellow (05)	○	○			

(102) Delineator Present
- ○ None (00)
- ○ Right (01)
- ○ Left (02)
- ○ Both Sides (03)

(103) Bikeway
- ○ None (00)
- ○ Bike Route [Signed] (01)
- ○ Bike Lane Stripe (02)
- ○ Separate Path/ Lane (03)

(104) Vehicle Factors (Select Up to 2)
- ○ None (00)
- ○ Worn Tires (01)
- ○ Tire Failure (02)
- ○ Brakes (03)
- ○ Headlights (04)
- ○ Taillights (05)
- ○ Signals (06)
- ○ Steering (07)
- ○ Suspension (08)
- ○ Wheels (09)
- ○ Power Train (10)
- ○ Window/Windshield (11)
- ○ Mirrors (12)
- ○ Wipers (13)
- ○ Trailer Coupling (14)
- ○ Other (15)

(105) Vision Obstruction (Select up to 2)
- ○ None (00)
- ○ Trees/Brush/Fence (01)
- ○ Embankment (02)
- ○ Building (03)
- ○ Moving Vehicle (04)
- ○ Parked/Stopped Vehicle (05)
- ○ Glare (06)
- ○ Weather Condition (07)
- ○ Pedestrian (08)
- ○ Animal(s) in Road (09)
- ○ Other (10)

(106) Human Factors (Select up to 2)
- ○ None (00)
- ○ Inattention (01)
- ○ Misjudgment (02)
- ○ Fatigue (03)
- ○ Alcohol (04)
- ○ Illegal Drugs (05)
- ○ Illness (06)
- ○ Legal Meds. (07)
- ○ Emotional (08)
- ○ Phys. Impaired (09)
- ○ Other (10)

(107) Driver Distracted By
- ○ Not Distracted (00)
- ○ Cellular Phone (01)
- ○ Other Elect. Comm. Device (02)
- ○ Other Electronic Device (03)
- ○ Other Inside Vehicle (04)
- ○ Other Outside Vehicle (05)
- ○ Other Occupant (06)

(108) Other Factors (Select up to 4)
- ○ No Improper Action (00)
- ○ Drove too Fast for Conditions (01)
- ○ Exceed Posted Speed Limit (02)
- ○ Disregard Traffic Signals (03)
- ○ Disregard Red Light (04)
- ○ Disregard Other Trfc. Ctrl. Dev. (05)
- ○ Failure to Yield (06)
- ○ Wrong Side/Way (07)
- ○ Crossed Centerline (08)
- ○ Ran Off Road (09)
- ○ Failure to Keep in Proper Lane (10)
- ○ Improper Turn (11)
- ○ Improper Passing (12)
- ○ Improper Backing (13)
- ○ Followed too Closely (14)
- ○ Aggressive, Reckless Driving (15)
- ○ Swerved to Avoid Obstacle (16)
- ○ Over Correcting or Over Steering (17)
- ○ Other Improper Action (18)
- ○ Illegally in Roadway (19)
- ○ Improper Crossing (20)
- ○ Pedestrian Viol. (21)
- ○ Inattention [Talking, Etc.] (22)
- ○ Bicycle Violation (23)
- ○ Clothing not Visible (24)

(109) Roadway Comp.
- ○ Concrete (01)
- ○ Asphalt (02)
- ○ Gravel (03)
- ○ Dirt (04)
- ○ Other (05)

(110) Roadway Surface
- ○ Dry (01)
- ○ Wet (02)
- ○ Mud, Dirt, Gravel (03)
- ○ Debris (04)
- ○ Oil (05)
- ○ Snow (06)
- ○ Slush (07)
- ○ Ice/Frost (08)
- ○ Water (09)
- ○ Sand (10)

(111) Other Roadway Conditions
- ○ None (00)
- ○ Ruts, Holes, Etc. (01)
- ○ No Shoulder (02)
- ○ Low Shoulder (03)
- ○ Soft Shoulder (04)
- ○ High Shoulder (05)
- ○ Loose Material (06)
- ○ Worn, Polished (07)
- ○ Other (08)

(112) Roadway Alignment (Horizontal)
- ○ Straight (01)
- ○ Curve Left (02)
- ○ Curve Right (03)

(113) Roadway Alignment (Vertical)
- ○ Level (01)
- ○ Hillcrest (02)
- ○ Uphill (03)
- ○ Downhill (04)
- ○ Sag (05)

Officer's Rank and Name	Officer's ID Number	Date/Time	Supervisor's Rank and Name	Supervisor's ID Number	Date/Time

Page _____ of _____
DOT-1-174D (HWY-T) Rev. 06/08

STATE OF HAWAII MOTOR VEHICLE ACCIDENT REPORT

Report Number:_____

DIAGRAM

(114) Tire Skid Marks (Feet)

Wheel	Unit	Unit	Unit	Unit
Rgt-R				
Lft-F				
Rgt-F				
Lft-R				

(115) REFERENCE POINT

IS _____ (feet) _____(direction)_____(Object/Landmark)

ALL OBJECTS ARE MEASURED FROM POINT OF REFERENCE

Object	N	S	E	W

(116) Intersection Related

○ No (01) ○ Yes (02)

(117) Main Road

(A) No. of Lanes	(B) Speed Limit

(118) Side Road

(A) No. of Lanes	(B) Speed Limit

(119) Indicate the Type of Intersection (Check one)

○ Not at Intersection (01) ○ "Y" Intersection (04) ○ Roundabout (07)
○ 4-Way Intersection (02) ○ Part of Interchange (05) ○ 5 (or more legs) Intersection (08)
○ "T" Intersection (03) ○ Traffic Circle (06) ○ Other (09)

Place an arrow in the above circle to indicate North.

Draw Object, Directions, Etc. According to Current Practices.

Synopsis (Accident Description. Refer to units by number):

Officer's Rank and Name	Officer's ID Number	Date/Time	Supervisor's Rank and Name	Supervisor's ID Number	Date/Time

Page _____ of _____
DOT-1-174E (HWY-T) Rev. 06/08

STATE OF HAWAII MOTOR VEHICLE ACCIDENT REPORT

Report Number:_____

(120) ALL PERSONS

E - Ejection
00 Not Ejected
01 Ejected, Total
02 Ejected, Partial
03 N/A Non-motorist
04 Unknown

F - Safety Equipment Use
00 Not Present
01 Not Used
02 Shoulder/Lap Belt Used
03 Lap Belt Only Used
04 Shoulder Belt Only Used
05 Not Able to Determine
06 Child Restraint (Forward)
07 Child Restraint (Rear)
08 Booster Seat
09 Child Restraint (Unk. Type)
10 Child Restraint (Improper)
11 Helmet Used
12 N/A (Non-Motorist)
13 Unknown

G - Air Bag Deployed
00 Not Present
01 Not Deployed
02 Deployed - Front
03 Deployed - Side
04 Deployed - Other
05 Deployed - Combination
06 Deployed - Curtain

H - Injury Class
00 None
01 Possible
02 Non-incapacitating
03 Incapacitating
04 Fatal
05 Unknown

I - Injury Area
00 None
01 Head
02 Face
03 Eye
04 Neck
05 Thorax (Chest)
06 Spine/Back
07 Shoulder/Upper Arm
08 Elbow/Lower Arm/Hand
09 Abdomen/Pelvis
10 Hip/Upper Leg
11 Knee/Lower Leg/Foot
12 Entire Body

J - Accident Site Care
00 None
01 First Aid
02 Resuscitation
03 Extrication
04 Both 1 & 2
05 Both 1 & 3
06 Both 2 & 3
07 Other
08 Refused

K - Trans. to Med. Facility
00 Not Transported
01 EMS
02 Police
03 Helicopter
04 Private Vehicle
05 Other

L - Medical Facility

Hawaii County
01 Hilo Medical Center
02 Kona Hospital
03 Kau Hospital
04 Kohala Hospital
05 Honokaa Hospital
06 N. Hawaii Comm. Hosp.

Maui County
07 Kula General Hospital
08 Maui Mem. Med. Ctr.
09 Kaiser Clinic
10 Hana Clinic

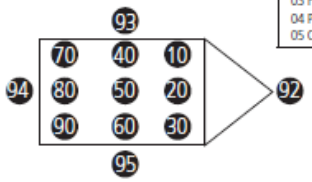
Molokai/Lanai
11 Molokai General Hosp.
12 Lanai Comm. Hospital

Kauai County
13 Wilcox Memorial Hosp.
14 Kauai Vet. Mem. Hosp.

C&C Honolulu
15 Castle Medical Center
16 Shriner's Hosp. for Children
17 Kahuku Hospital
18 Kaiser Permanente
19 Kaiser Clinic - Honolulu

C&C Honolulu (cont.)
20 Kaneohe State Hospital
21 Kapiolani Medical Ctr.
22 Kapiolani Med. - Pali Momi
23 Kuakini Med. Ctr.
24 Hawaii Med. Ctr.
25 Hawaii Med. Ctr. West
26 Queen's Medical Center
27 Straub Clinic & Hosp.
28 Tripler Army Med. Ctr.
29 Wahiawa General Hosp.
30 Waianae Comp. Ctr.

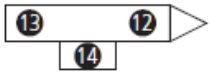
99 Other



Motor Vehicle
For lap positions use 1 in place of 0

B - Position in Unit

Motorcycle/Moped/Bicycle



Pedestrian

15

M - Condition
01 Refused Treatment
02 Released
03 Good, Fair
04 Serious, Guarded
05 Critical
06 Dead on Arrival
07 Dead Other

Name and Address	A Unit	B Posit.	C Age	D Sex	E Eject	F Safety	G Air Bag	H Injury	I Area	J Care	K Trans	L Hosp.	M Cond.	N EMS No.

Officer's Rank and Name	Officer's ID Number	Date/Time	Supervisor's Rank and Name	Supervisor's ID Number	Date/Time

B. Summary of statistics

Table 8 presents the descriptive statistics of the continuous explanatory variables for the 0.1-mile segment length.

Table 8 Descriptive statistics of the continuous variables

Variable	Min	Max	Mean	S.D.
AADT	545	26682	6788	5906
The Proportion of Single and Combination Trucks in the Stream	0.00	56.57	5.60	7.32
Average Side Friction Demand	0.00	0.40	0.02	0.03
Absolute Value of Curvature(degrees)	0.00	64.79	3.43	7.03
Standard Deviation of Curvature	0.00	70.26	2.84	6.73
Absolute Value of Curvature (Environmental Variable)	0.00	35.57	3.38	5.95
Standard Deviation of Curvature (Environmental Variable)	0.00	41.46	4.57	7.15
Grade (percent)	-16.30	12.32	0.00	3.16
Standard Deviation of Grade	0.00	7.32	0.27	0.52
Grade (Environmental Variable)	0.00	10.95	2.40	1.78
Standard Deviation of Grade (Environmental Variable)	0.00	7.40	1.30	0.94
IRI (inch/mile)	0.00	768.30	145.68	74.65
Standard Deviation of IRI	0.00	566.27	42.21	36.15
IRI (Environmental Variable)	34.40	591.18	145.01	65.46
Standard Deviation of IRI (Environmental Variable)	6.41	566.27	55.63	35.36
Lane Width (feet)	7.00	19.40	10.99	1.26
Painted Median Width (feet)	0.00	12.00	0.16	1.28
Rutting (inch)	0.00	0.82	0.06	0.05
Shoulder Width (feet)	0.00	20.00	4.81	3.09
The Proportion of Total Length of sections with Guardrails(mile/mile)	0.00	1.00	0.20	0.37
The Proportion of Total Length of Sections with painted Medians (mile/mile)	0.00	1.00	0.08	0.25
The Proportion of Total Length of Sections with Asphalt Concrete Shoulder(mile/mile)	0.00	1.00	0.80	0.40

Table 9 presents the descriptive statistics of the continuous explanatory variables for the 0.3-mile segment length.

Table 9 Descriptive statistics of the continuous variables

Variable	Min	Max	Mean	S.D.
AADT	545	26682	6785	5881
The Proportion of Single and Combination Trucks in the Stream	0.00	56.57	5.59	7.28
Average Side Friction Demand	0.00	0.26	0.03	0.03
Absolute Value of Curvature(degrees)	0.00	49.39	3.42	6.48
Standard Deviation of Curvature	0.00	53.33	3.95	7.34
Absolute Value of Curvature (Environmental Variable)	0.00	35.57	3.36	5.95
Standard Deviation of Curvature (Environmental Variable)	0.00	41.46	4.57	7.15
Grade (percent)	-12.63	11.99	0.00	3.06
Standard Deviation of Grade	0.00	7.28	0.66	0.73
Grade (Environmental Variable)	0.00	10.81	2.40	1.78
Standard Deviation of Grade (Environmental Variable)	0.00	7.38	1.30	0.94
IRI (inch/mile)	32.20	659.17	145.49	69.93
Standard Deviation of IRI	4.59	336.34	49.54	36.16
IRI (Environmental Variable)	34.48	591.18	144.54	65.12
Standard Deviation of IRI (Environmental Variable)	6.41	336.34	55.45	34.79
Lane Width (feet)	7.00	15.90	11.00	1.25
Painted Median Width (feet)	0.00	12.00	0.16	1.17
Rutting (inch)	0.00	0.53	0.06	0.05
Shoulder Width (feet)	0.00	16.00	4.81	3.01
The Proportion of Total Length of sections with Guardrails (mile/mile)	0.00	1.00	0.20	0.32
The Proportion of Total Length of Sections with painted Medians (mile/mile)	0.00	1.00	0.08	0.22
The Proportion of Total Length of Sections with Asphalt Concrete Shoulder (mile/mile)	0.00	1.00	0.81	0.42

Table 10 presents the descriptive statistics of the continuous explanatory variables for the 0.5-mile segment length.

Table 10 Descriptive statistics of the continuous variables

Variable	Min	Max	Mean	S.D.
AADT	545	26682	6796	5882
The Proportion of Single and Combination Trucks in the Stream	0.00	56.57	5.61	7.29
Average Side Friction Demand	0.00	0.36	0.03	0.03
Absolute Value of Curvature(degrees)	0.00	47.53	3.42	6.30
Standard Deviation of Curvature	0.00	51.38	4.26	7.34
Absolute Value of Curvature (Environmental Variable)	0.00	36.56	3.33	5.81
Standard Deviation of Curvature (Environmental Variable)	0.00	41.70	4.69	7.04
Grade (percent)	-11.21	11.20	0.00	2.96
Standard Deviation of Grade	0.00	8.57	0.91	0.85
Grade (Environmental Variable)	0.01	8.45	2.40	1.70
Standard Deviation of Grade (Environmental Variable)	0.00	6.39	1.49	0.94
IRI (inch/mile)	32.90	656.30	145.48	68.59
Standard Deviation of IRI	6.18	265.15	51.95	35.20
IRI (Environmental Variable)	35.70	568.10	143.90	62.97
Standard Deviation of IRI (Environmental Variable)	7.52	265.15	57.16	33.93
Lane Width (feet)	7.00	14.34	11.00	1.24
Painted Median Width (feet)	0.00	12.00	0.16	1.00
Rutting (inch)	0.00	0.46	0.06	0.04
Shoulder Width (feet)	0.00	16.00	4.81	2.97
The Proportion of Total Length of sections with Guardrails (mile/mile)	0.00	1.00	0.20	0.30
The Proportion of Total Length of Sections with painted Medians (mile/mile)	0.00	1.00	0.08	0.20
The Proportion of Total Length of Sections with Asphalt Concrete Shoulder (mile/mile)	0.00	1.00	0.81	0.41

C. Results for the other segment lengths

C.1 Negative binomial evaluation for 0.1-mile segment length

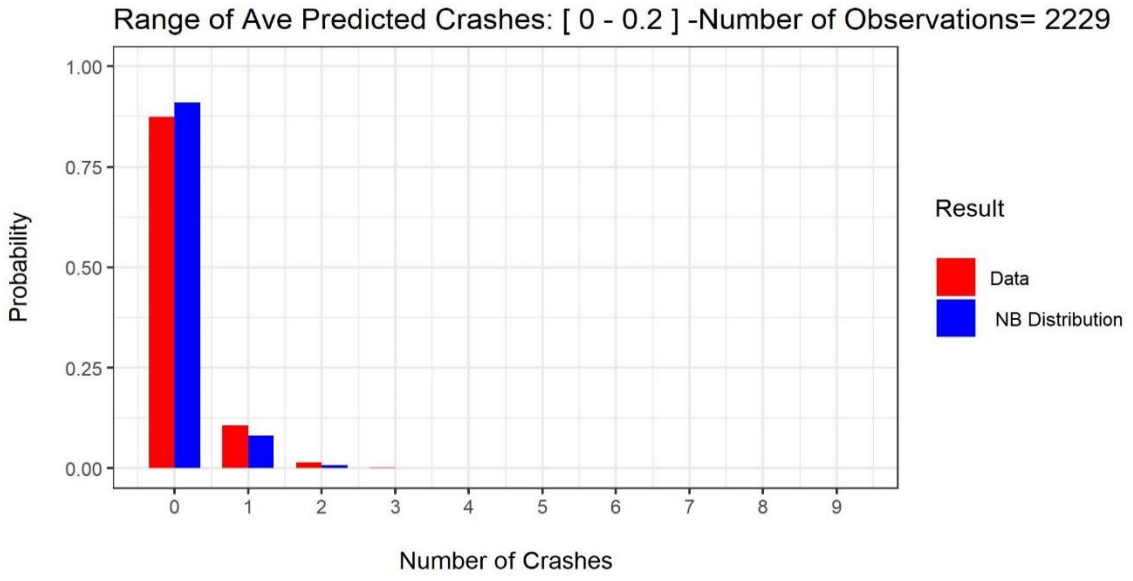


Figure 62 Comparison graph for the first range of average predicted crashes

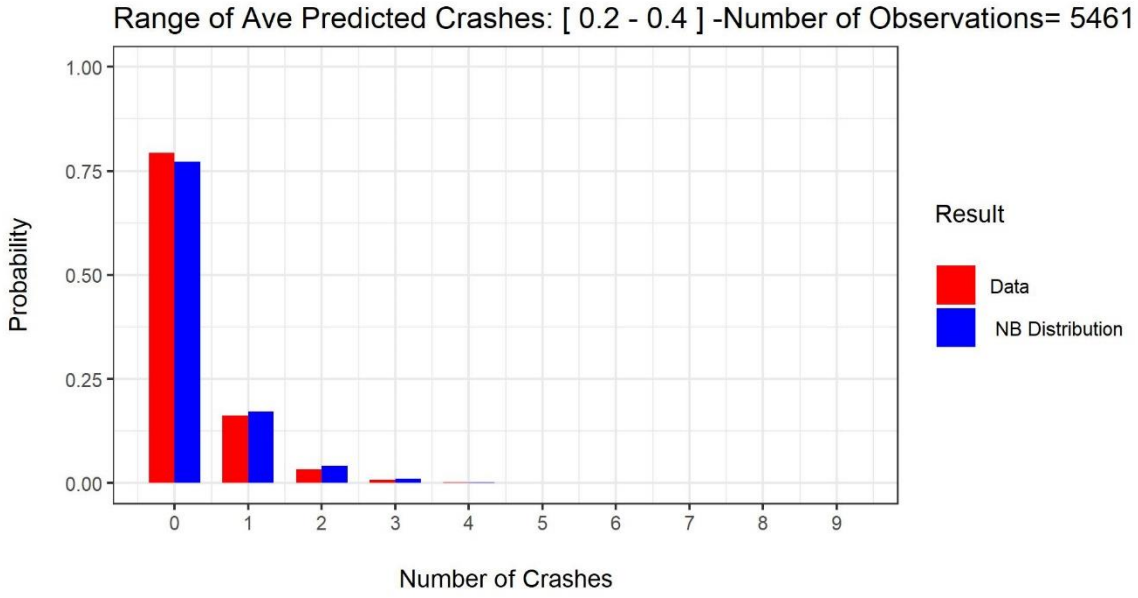


Figure 63 Comparison graph for the second range of average predicted crashes

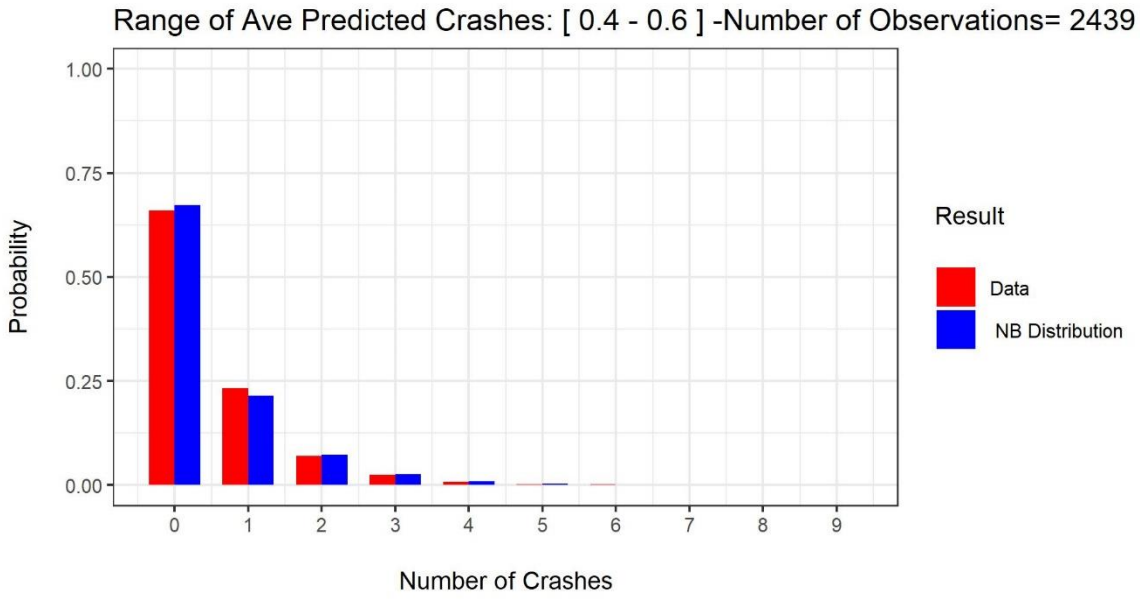


Figure 64 Comparison graph for the third range of average predicted crashes

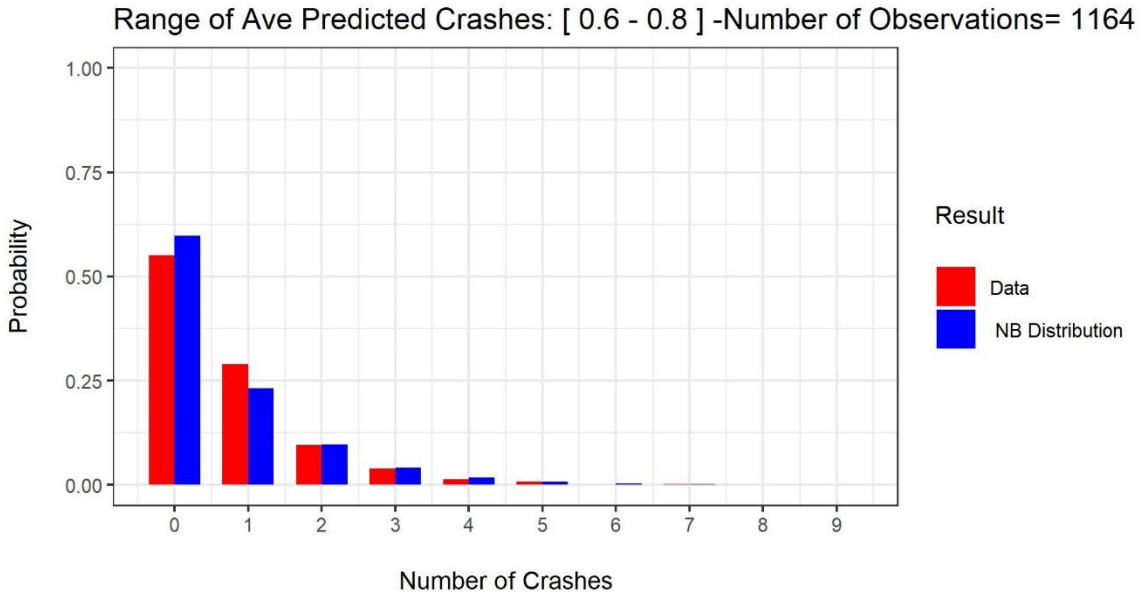


Figure 65 Comparison graph for the fourth range of average predicted crashes

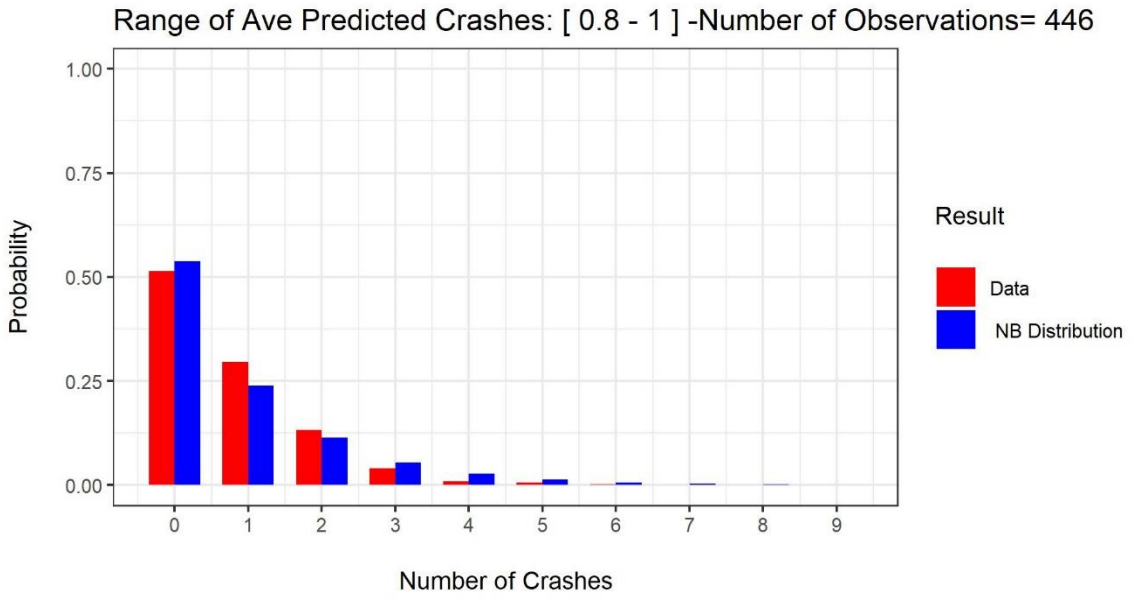


Figure 66 Comparison graph for the fifth range of average predicted crashes

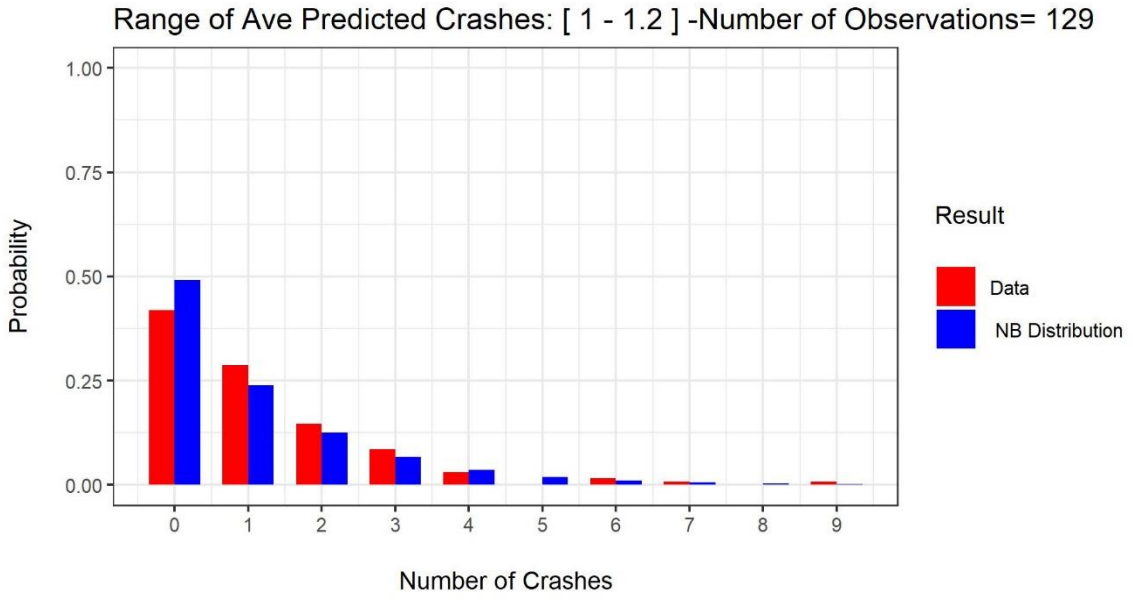


Figure 67 Comparison graph for the sixth range of average predicted crashes

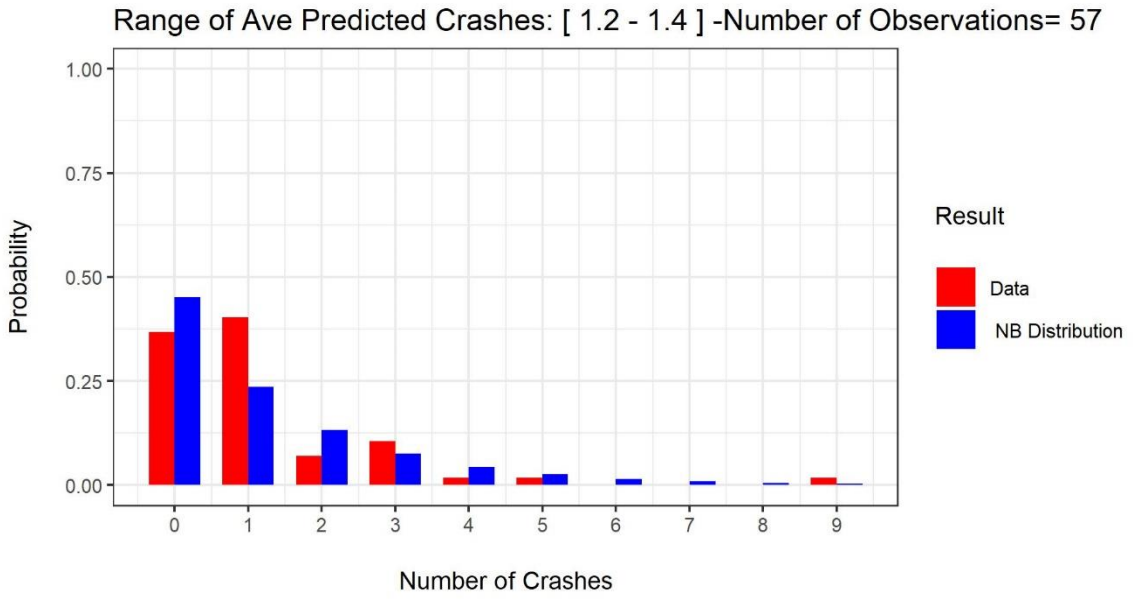


Figure 68 Comparison graph for the seventh range of average predicted crashes

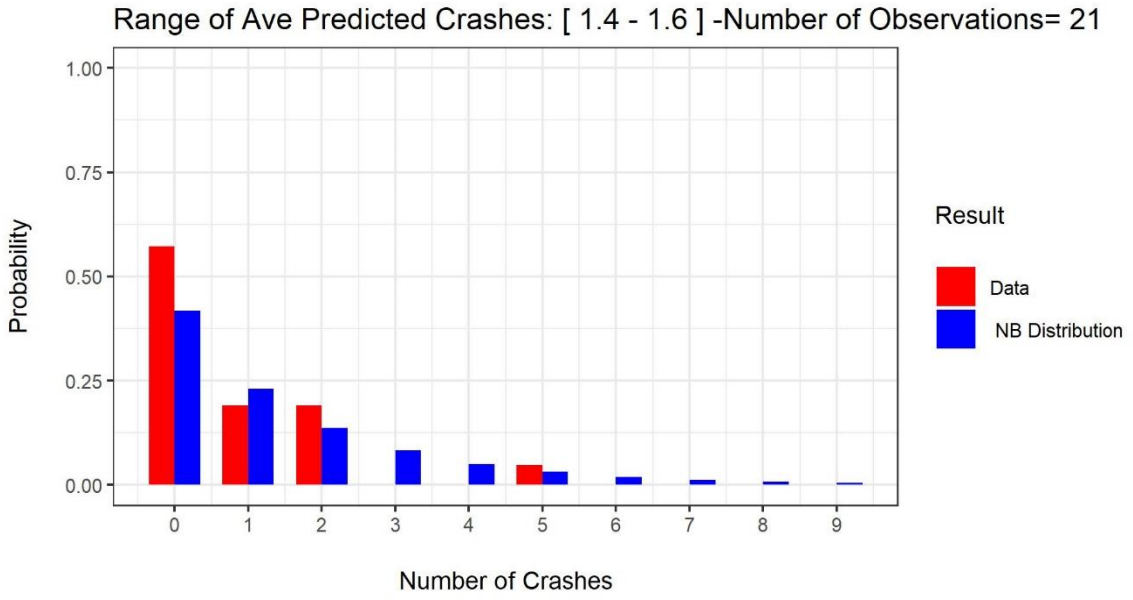


Figure 69 Comparison graph for the eighth range of average predicted crashes

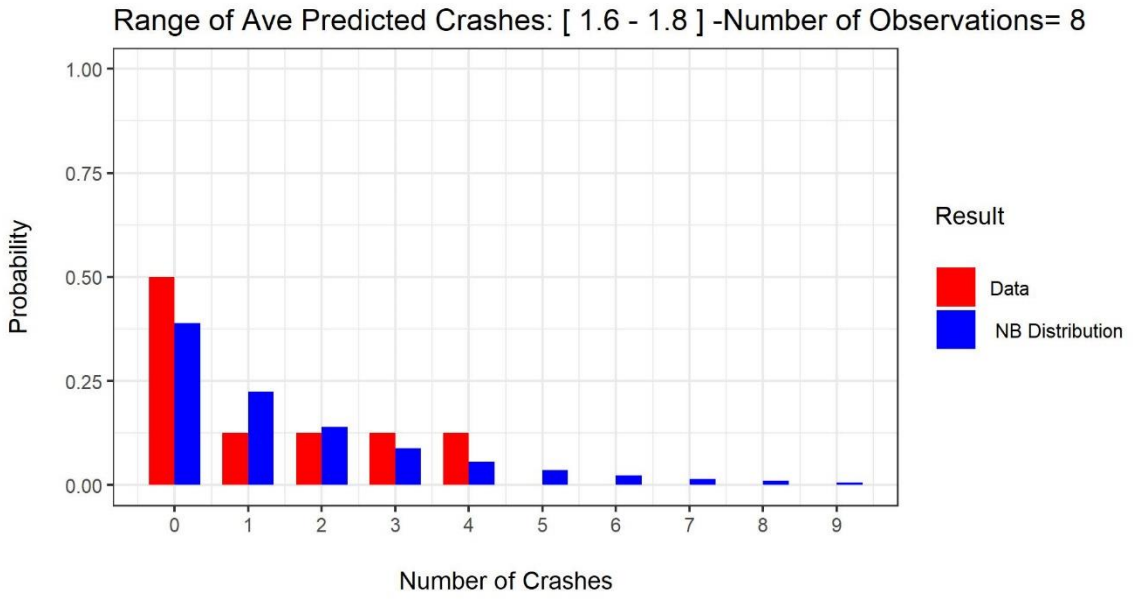


Figure 70 Comparison graph for the ninth range of average predicted crashes

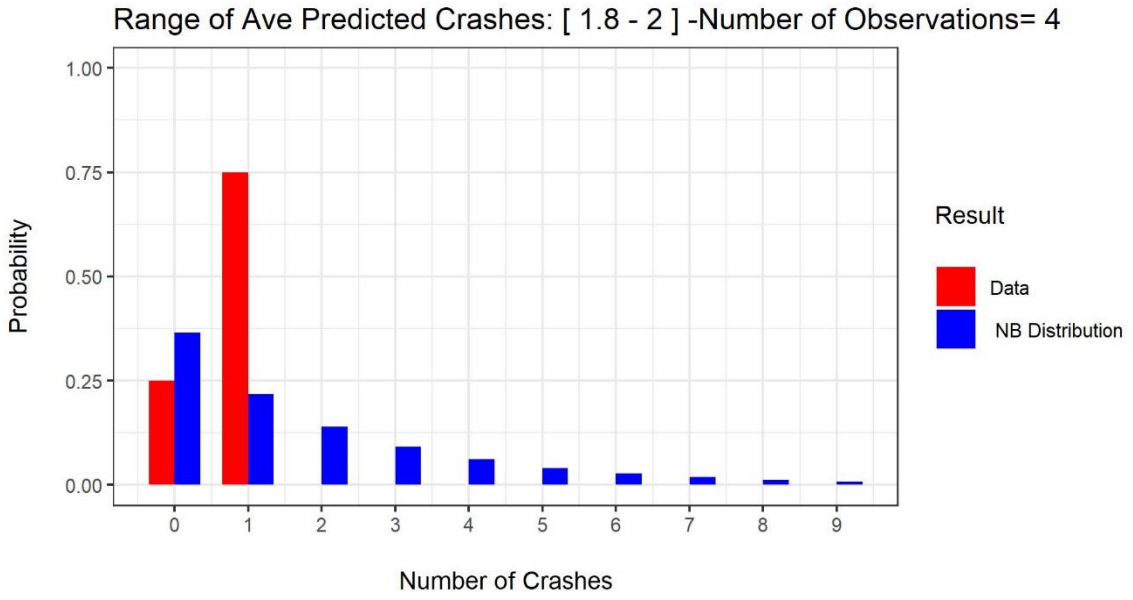


Figure 71 Comparison graph for the tenth range of average predicted crashes

C.2 Negative binomial evaluation for 0.3-mile segment length

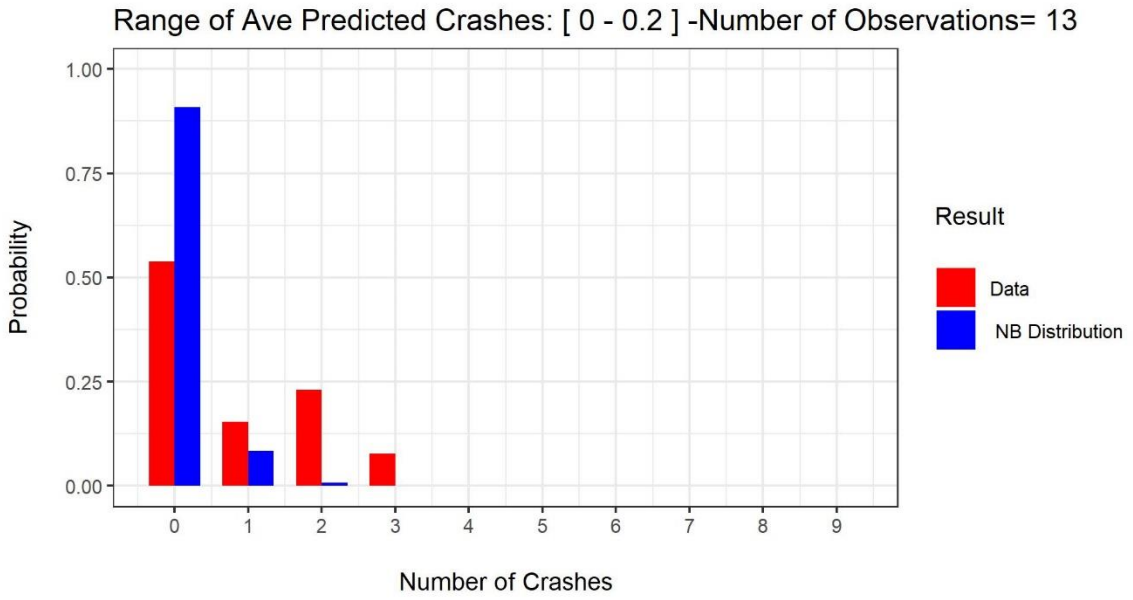


Figure 72 Comparison graph for the first range of average predicted crashes

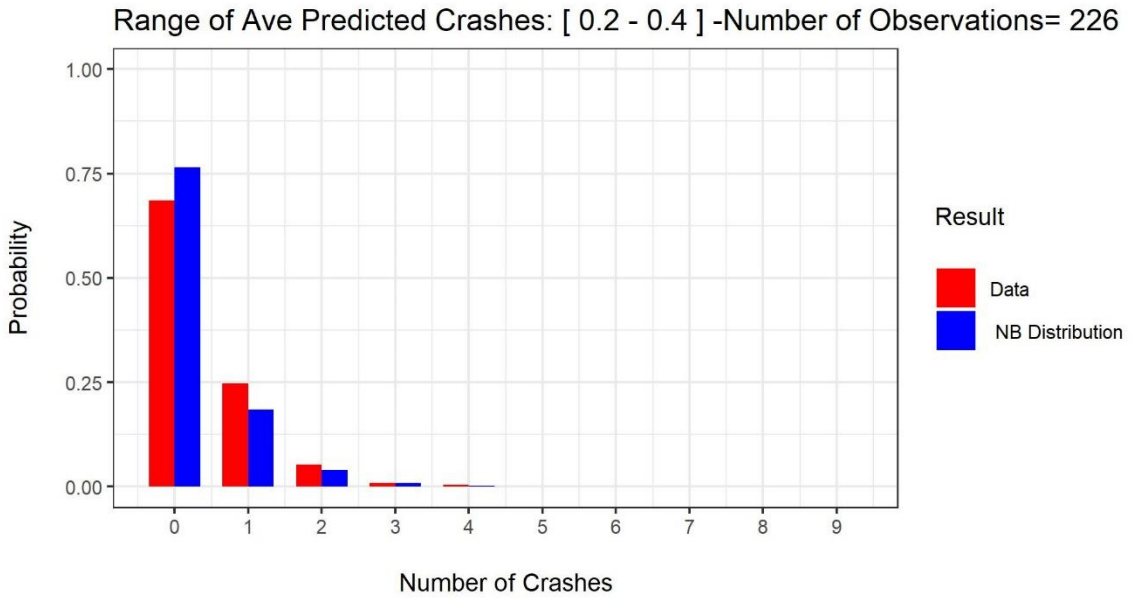


Figure 73 Comparison graph for the second range of average predicted crashes

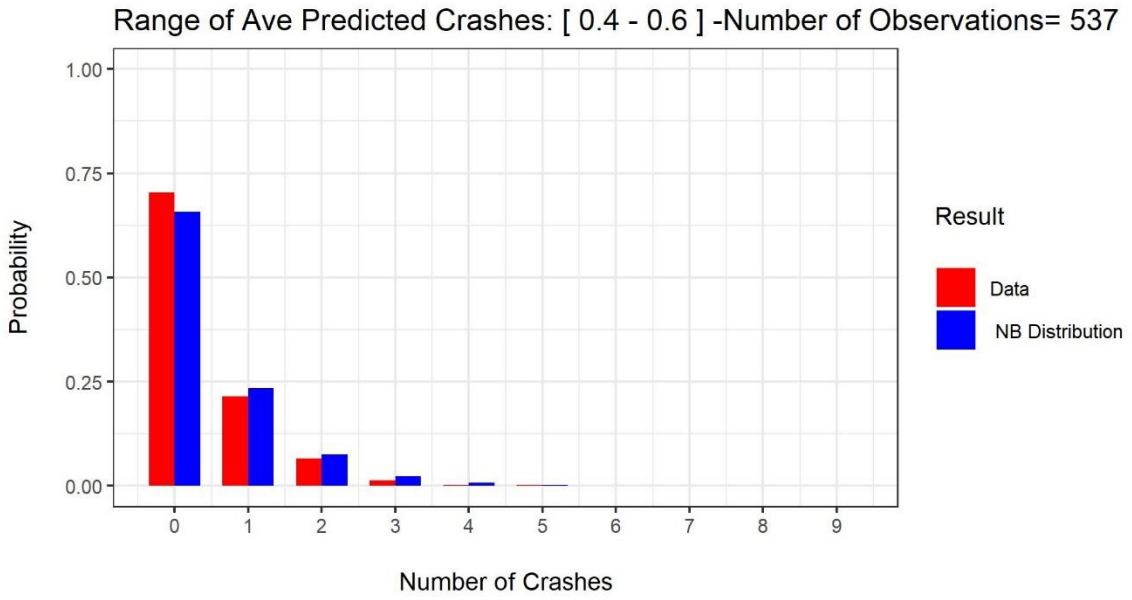


Figure 74 Comparison graph for the third range of average predicted crashes

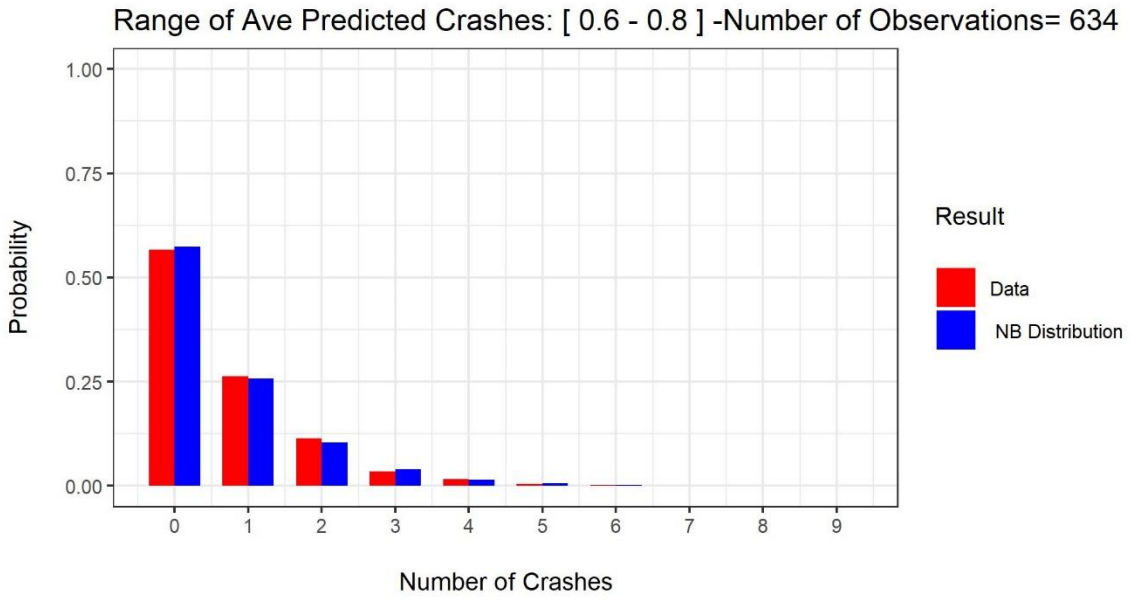


Figure 75 Comparison graph for the fourth range of average predicted crashes

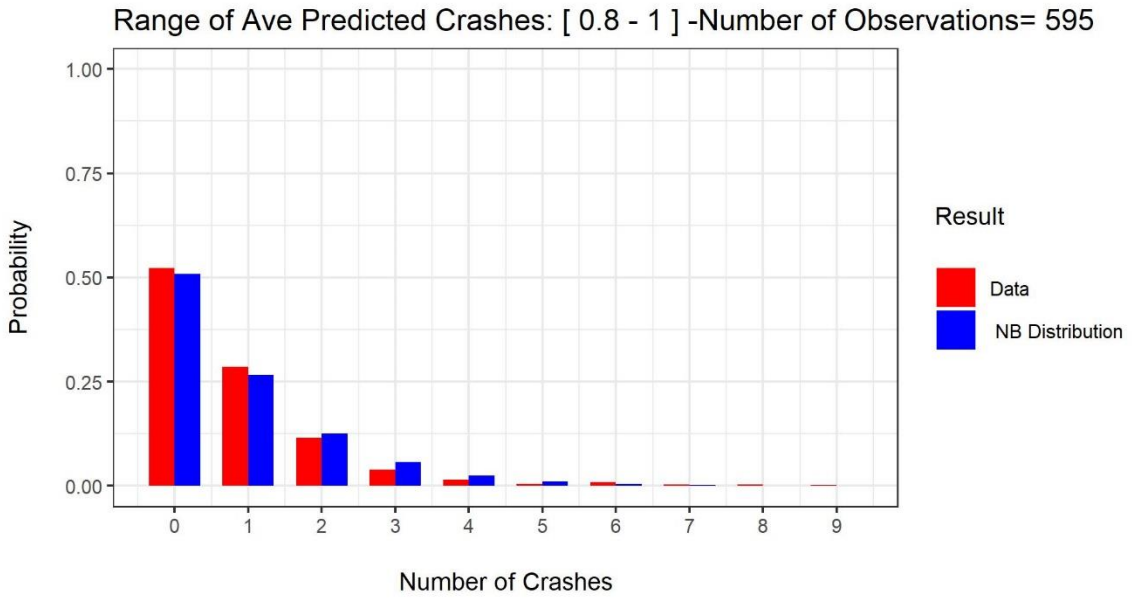


Figure 76 Comparison graph for the fifth range of average predicted crashes

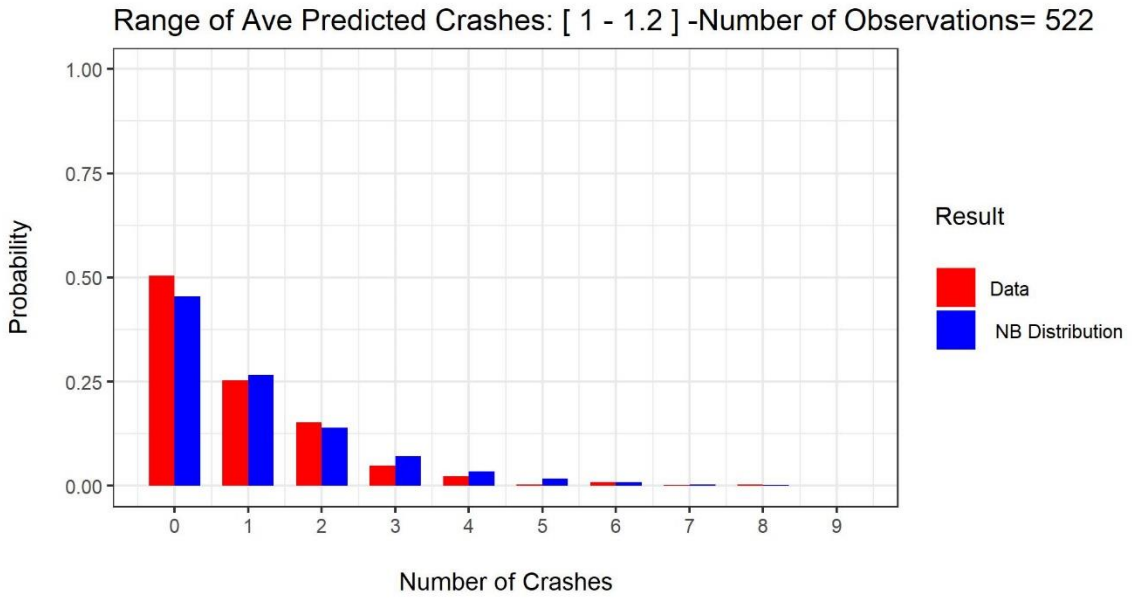


Figure 77 Comparison graph for the sixth range of average predicted crashes

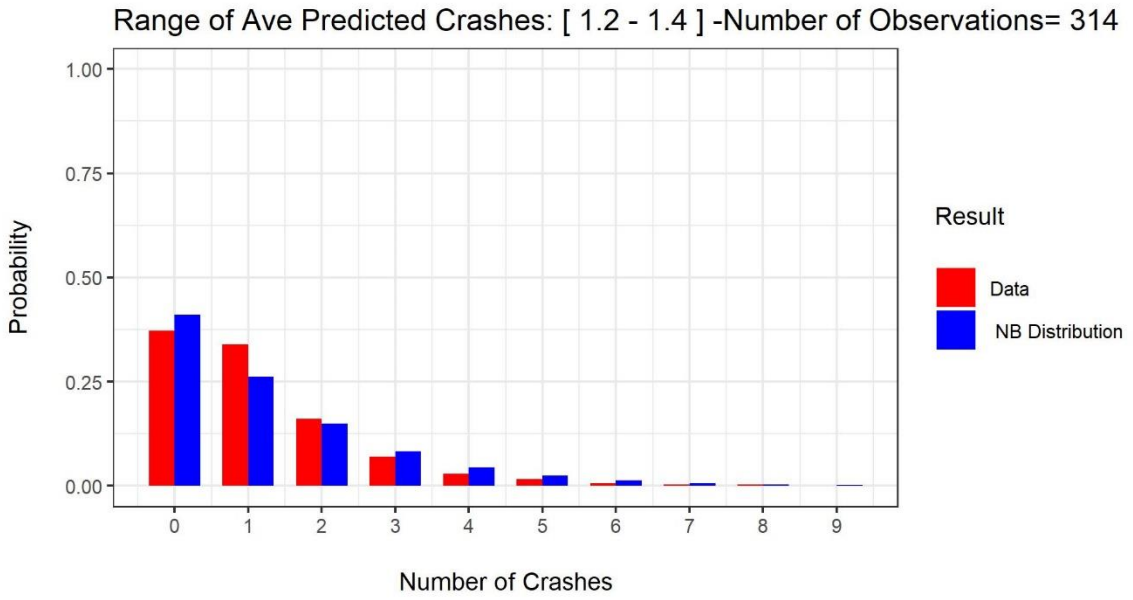


Figure 78 Comparison graph for the seventh range of average predicted crashes

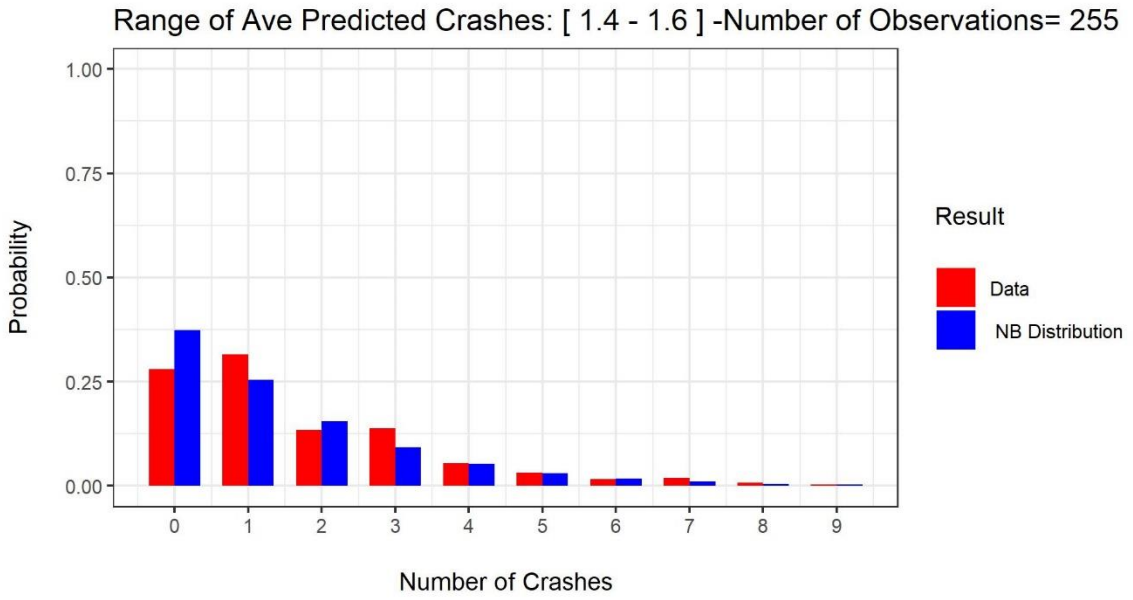


Figure 79 Comparison graph for the eighth range of average predicted crashes

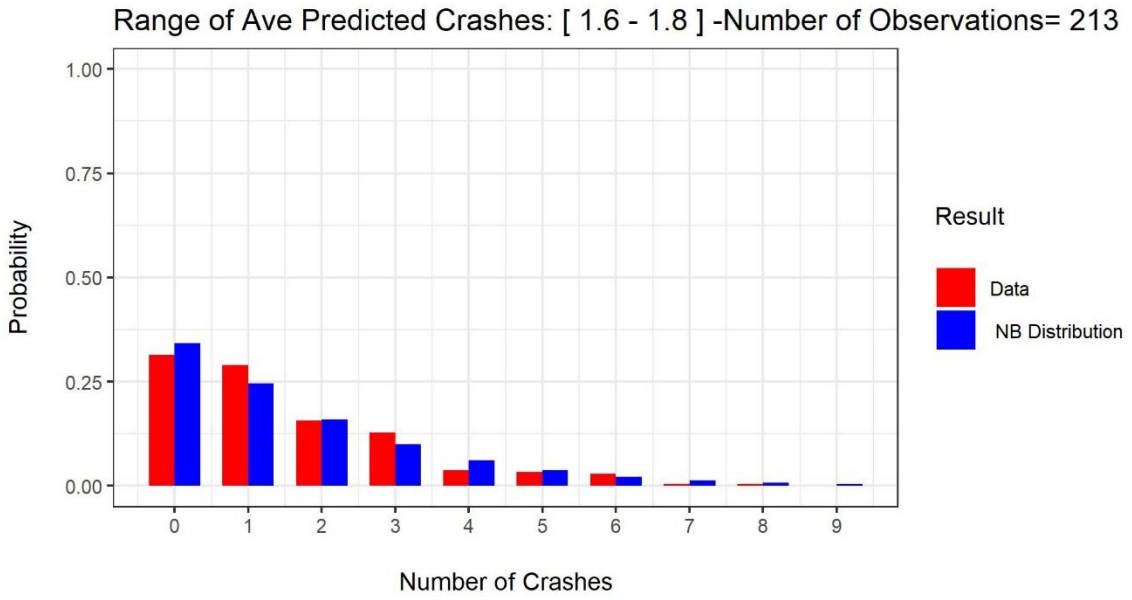


Figure 80 Comparison graph for the ninth range of average predicted crashes

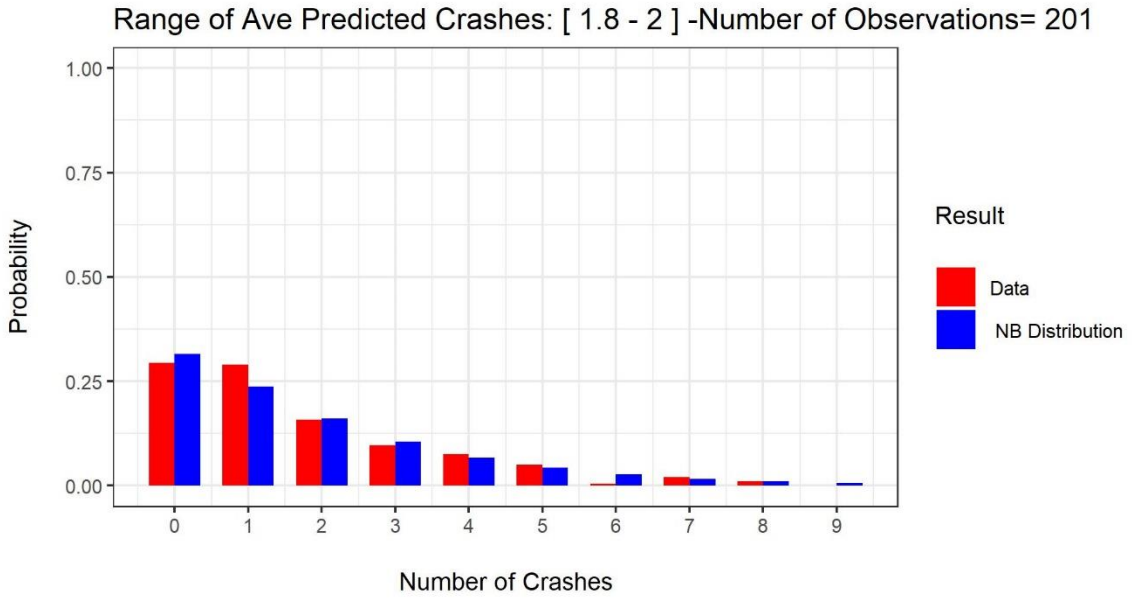


Figure 81 Comparison graph for the tenth range of average predicted crashes

C.3 Negative binomial evaluation for 0.5-mile segment length

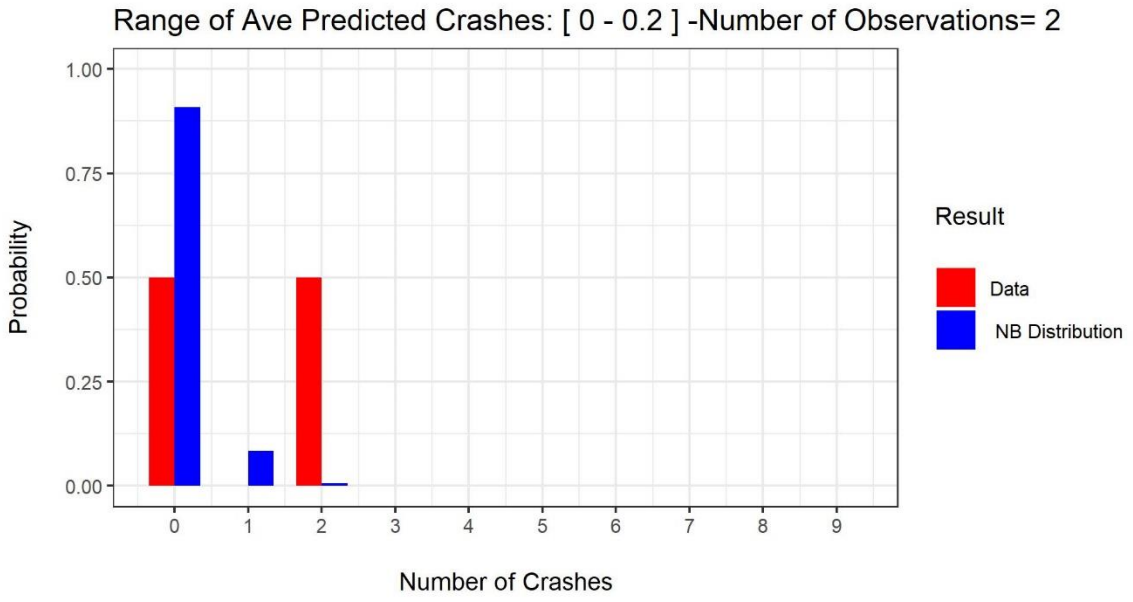


Figure 82 Comparison graph for the first range of average predicted crashes

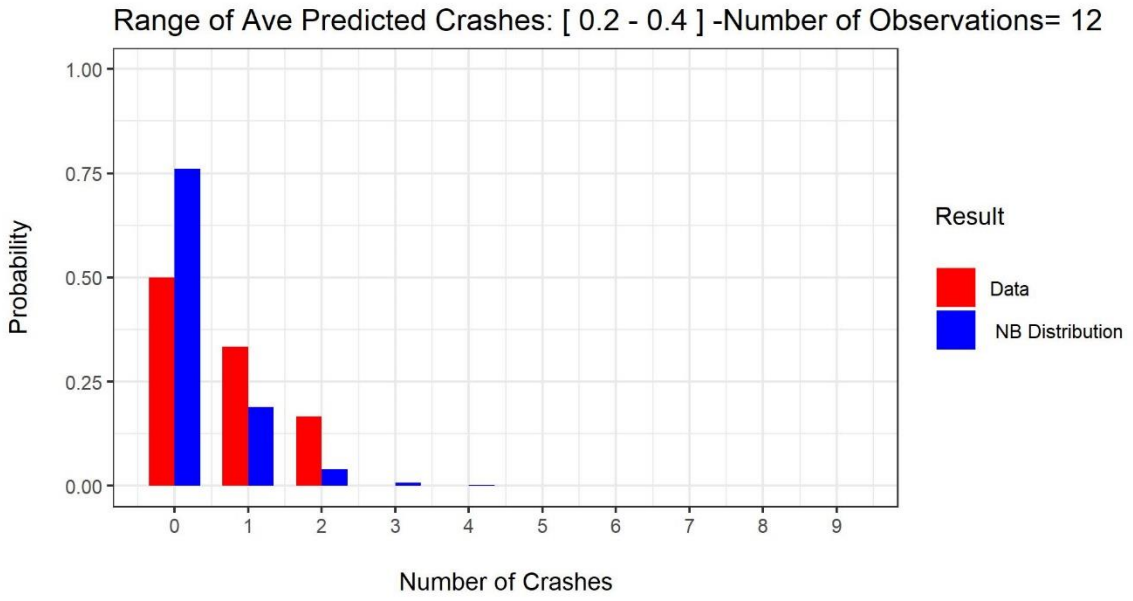


Figure 83 Comparison graph for the second range of average predicted crashes

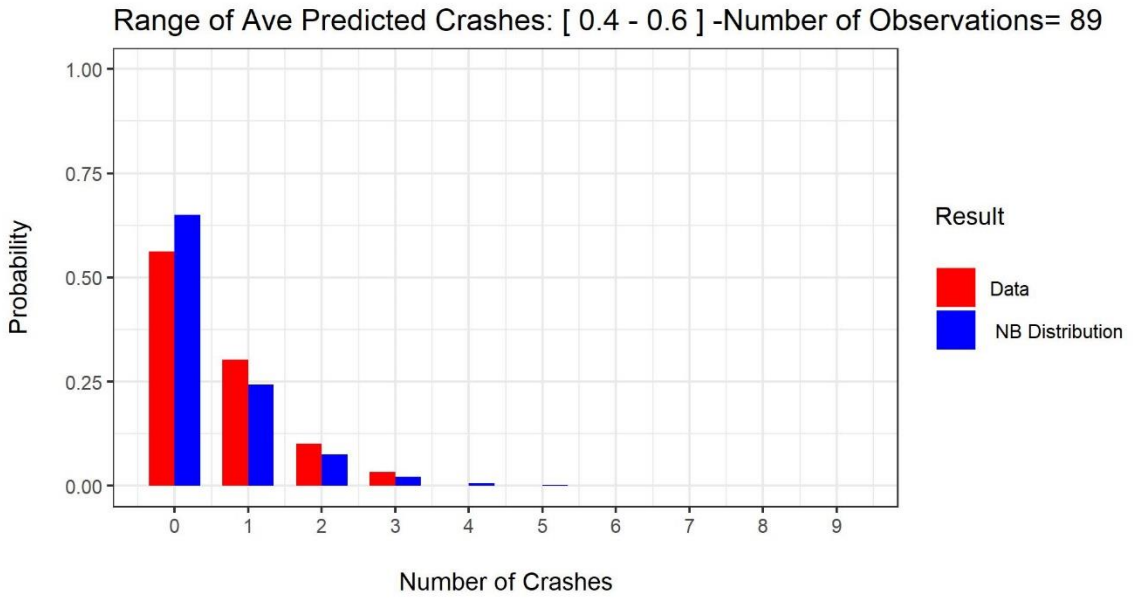


Figure 84 Comparison graph for the third range of average predicted crashes

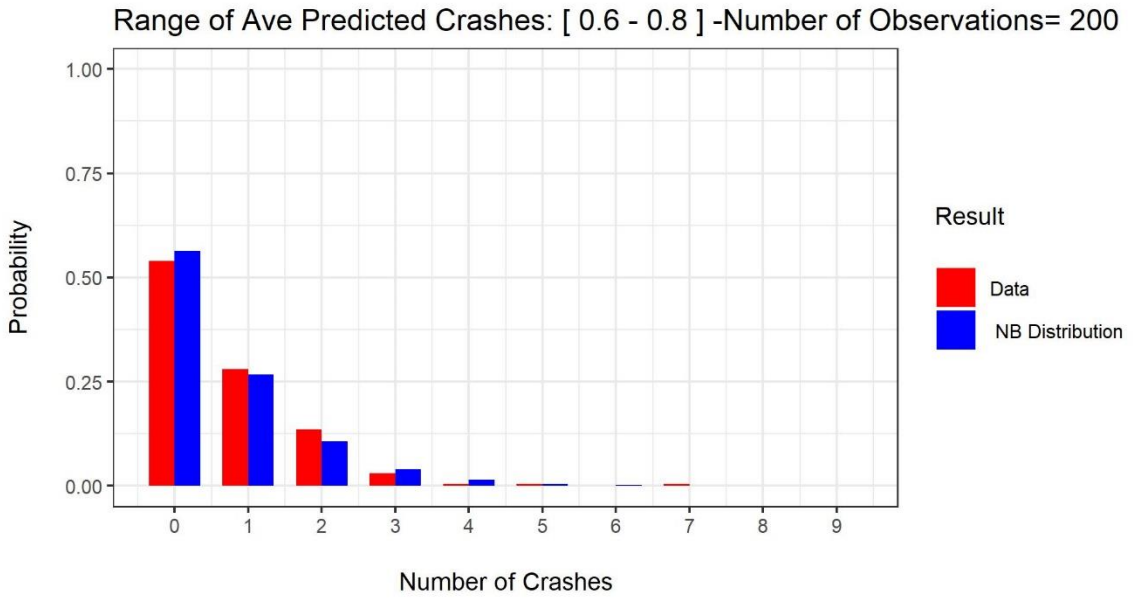


Figure 85 Comparison graph for the fourth range of average predicted crashes

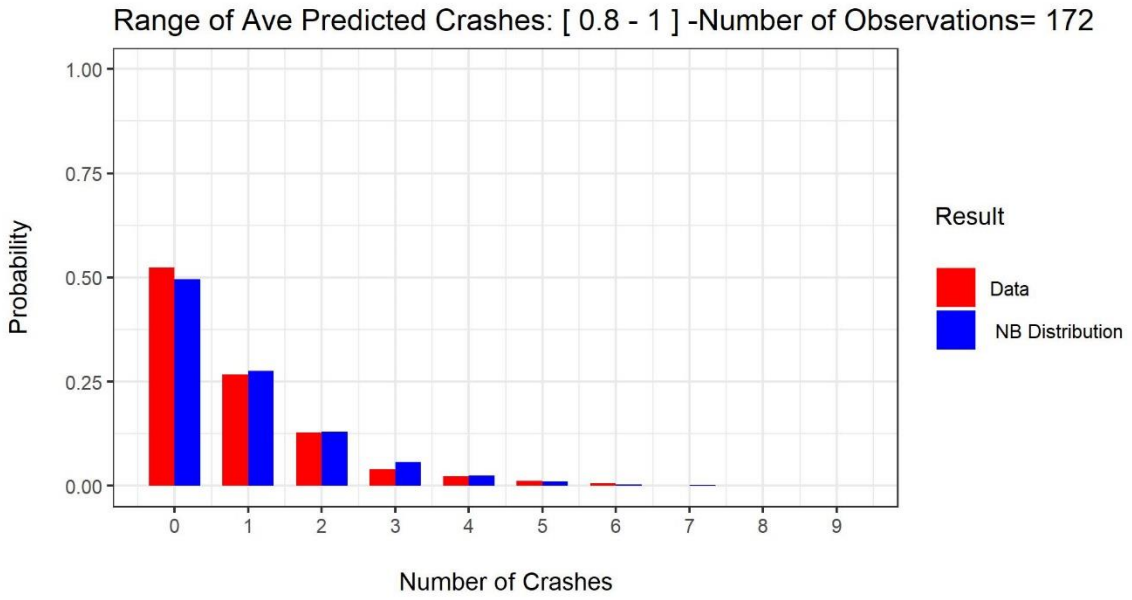


Figure 86 Comparison graph for the fifth range of average predicted crashes

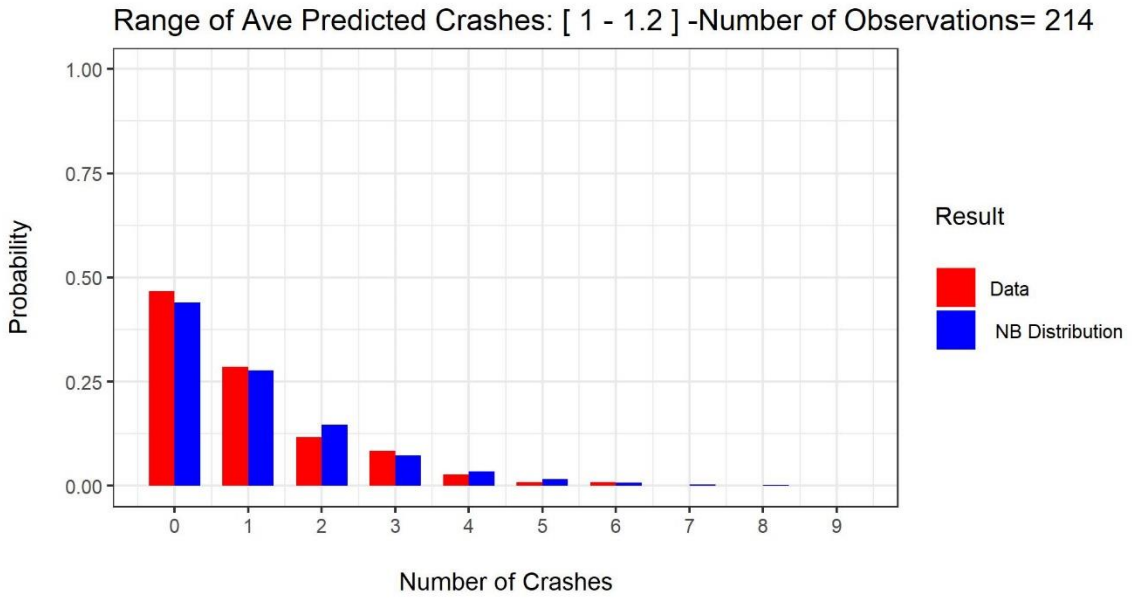


Figure 87 Comparison graph for the sixth range of average predicted crashes

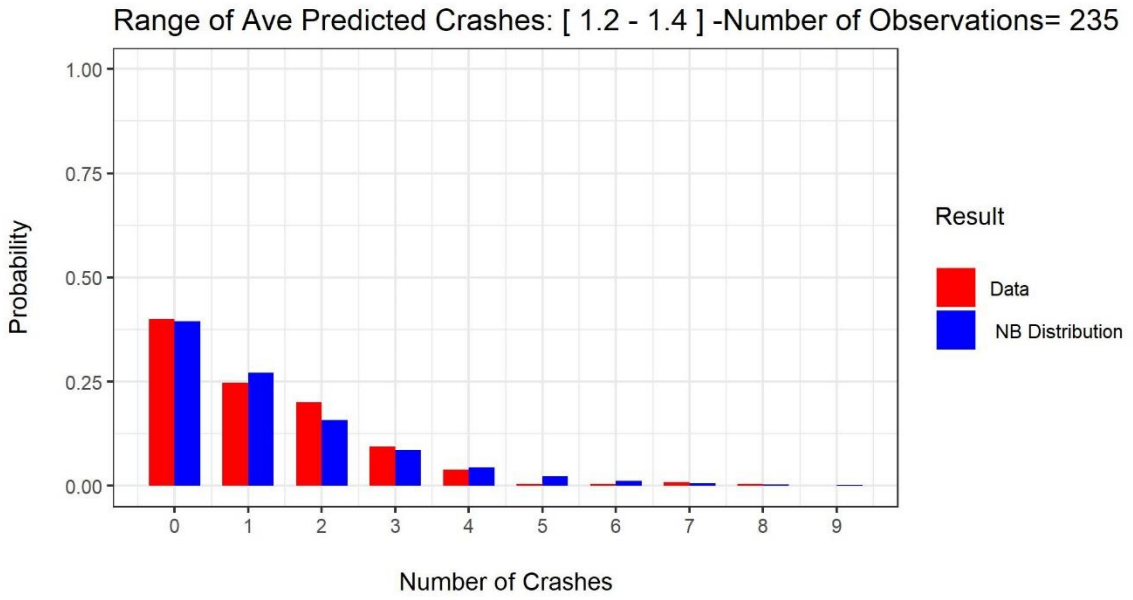


Figure 88 Comparison graph for the seventh range of average predicted crashes

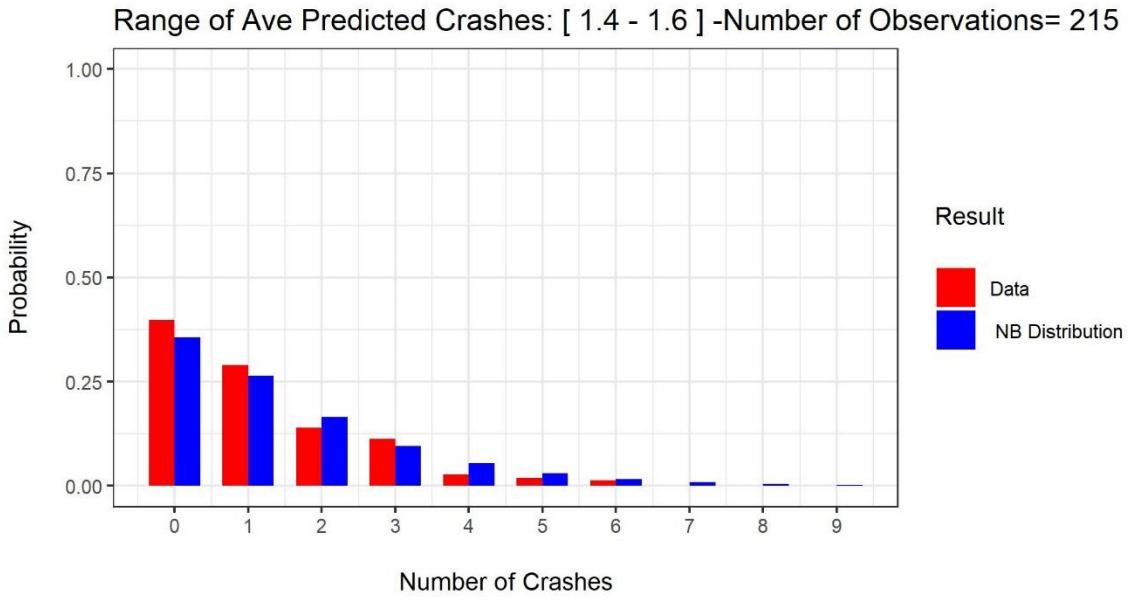


Figure 89 Comparison graph for the eighth range of average predicted crashes

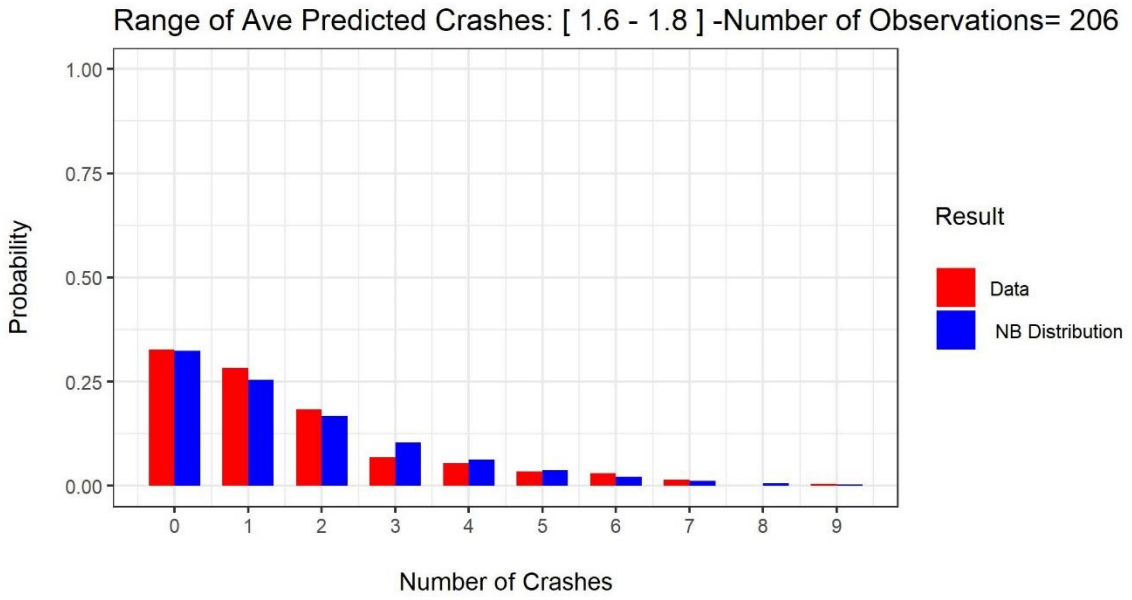


Figure 90 Comparison graph for the ninth range of average predicted crashes

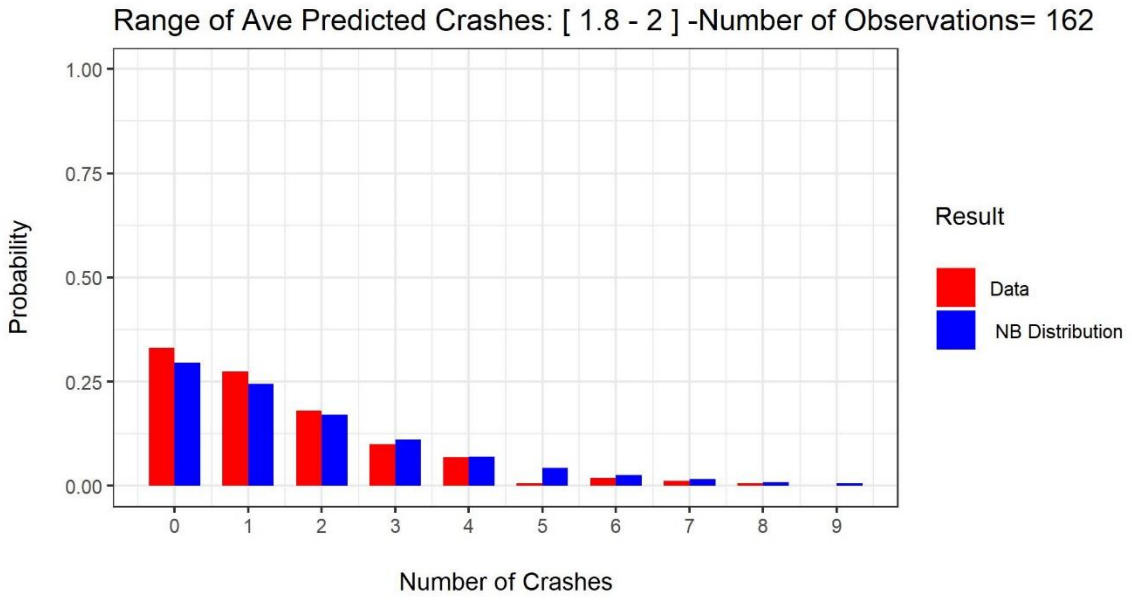


Figure 91 Comparison graph for the tenth range of average predicted crashes